

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

#2
T.D.
11/13/02

In re U.S. Patent Application of)
)
FUJIMOTO)
)
Application Number: -To be Assigned)
)
Filed: Concurrently Herewith)
)
For: STORAGE SYSTEM, DISK CONTROL CLUSTER AND)
A METHOD OF INCREASING OF DISK CONTROL)
CLUSTER)

Jc879 U.S. PTO
10/067332
02/07/02

Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

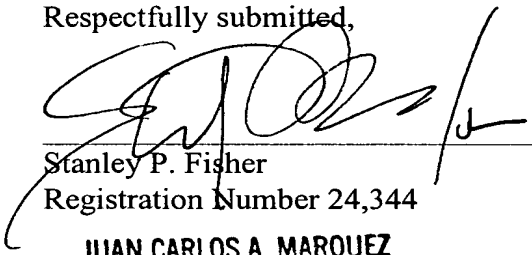
**REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION**

Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of September 26, 2001, the filing date of the corresponding Japanese patent application 2001-294048.

The certified copy of corresponding Japanese patent application 2001-294048 is being submitted herewith. Acknowledgment of receipt of the certified copies is respectfully requested in due course.

Respectfully submitted,


Stanley P. Fisher
Registration Number 24,344

JUAN CARLOS A. MARQUEZ
Registration No. 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
February 7, 2002

日本国特許庁
JAPAN PATENT OFFICE

879 U.S. PTO
10/067332
20/07/02

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日

Date of Application:

2001年 9月26日

出願番号

Application Number:

特願2001-294048

出願人

Applicant(s):

株式会社日立製作所

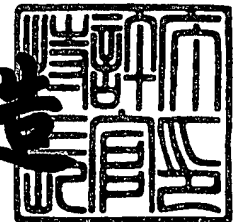
CERTIFIED COPY OF
PRIORITY DOCUMENT

Best Available Copy

2001年12月 7日

特許庁長官
Commissioner,
Japan Patent Office

及川耕造



出証番号 出証特2001-3106814

【書類名】 特許願

【整理番号】 H101260I

【提出日】 平成13年 9月26日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 12/00

【発明者】

【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社
日立製作所 中央研究所内

【氏名】 藤本 和久

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100099298

【弁理士】

【氏名又は名称】 伊藤 修

【連絡先】 0 3 - 3 2 5 1 - 3 8 2 4

【選任した代理人】

【識別番号】 100099302

【弁理士】

【氏名又は名称】 笹岡 茂

【手数料の表示】

【予納台帳番号】 018647

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 ストレージシステム、ディスク制御クラスタおよびディスク制御クラスタの増設方法

【特許請求の範囲】

【請求項1】 ホストコンピュータとのインターフェースを有する1または複数のチャネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報及び前記ディスク装置の管理情報を格納するローカル共有メモリ部とを有し、前記チャネルインターフェース部は前記ホストコンピュータからのデータのリード／ライト要求に対し、前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は、前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、

前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであって、

該各ディスク制御クラスタと他の各ディスク制御クラスタは相互結合網に接続され、該相互結合網に前記グローバル共有メモリ部が接続されたことを特徴とするストレージシステム。

【請求項2】 請求項1記載のストレージシステムにおいて、

前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部と前記ローカル共有メモリ部とが接続された接続部と他の各ディスク制御クラスタ内の該接続部が前記相互結合網を介して接続されたことを特徴とするストレージシステム。

【請求項3】 請求項1記載のストレージシステムにおいて、

前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部に前記ディスク制御クラスタ内で直接接続され、該各ディスク制御クラスタ内の該ローカル共有メモリ部と他の各ディスク制御クラスタ内の該ローカル共有メモリ部が前記相互結合網を介

して接続されたことを特徴とするストレージシステム。

【請求項 4】 請求項 1 記載のストレージシステムにおいて、

前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部に前記ディスク制御クラスタ内で直接接続され、

該各ディスク制御クラスタ内の該チャンネルインターフェース部と前記ディスクインターフェース部との接続部と他の各ディスク制御クラスタ内の該接続部が前記相互結合網を介して接続されたことを特徴とするストレージシステム。

【請求項 5】 ホストコンピュータとのインターフェースを有する 1 または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する 1 または複数のディスクインターフェース部と、該 1 または複数のチャンネルインターフェース部と該 1 または複数のディスクインターフェース部を接続する接続部を有する複数のディスク制御クラスタと、

前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報、前記ディスク装置の管理情報、及び前記ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであって、

前記各ディスク制御クラスタ内の接続部と他の各ディスク制御クラスタ内の該接続部が相互結合網を介して接続され、該相互結合網に前記グローバル共有メモリ部が接続されたことを特徴とするストレージシステム。

【請求項 6】 ホストコンピュータとのインターフェースを有する 1 または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する 1 または複数のディスクインターフェース部と、

前記ディスク装置に対しリード／ライトされるデータを格納する第 1 のメモリと、前記チャンネルインターフェース部及び前記ディスクインターフェース部と前記第 1 のメモリとの間のデータ転送に関する制御情報及び前記ディスク装置の管理情報を格納する第 2 のメモリとを有するローカル共有メモリ部とを有し、前記チャンネルインターフェース部は前記ホストコンピュータからのデータのリード／ライト要求に対し、前記ホストコンピュータとのインターフェースと前記ローカル

共有メモリ部内の前記第 1 のメモリとの間のデータ転送を実行し、前記ディスクインターフェース部は、前記ディスク装置と前記ローカル共有メモリ部内の前記第 1 のメモリとの間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、

前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであって、

前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部の第 2 のメモリに前記ディスク制御クラスタ内で直接接続され、

該各ディスク制御クラスタ内の該チャンネルインターフェース部と前記ディスクインターフェース部との第 1 の接続部と他の各ディスク制御クラスタ内の該第 1 の接続部が第 1 の相互結合網を介して接続され、

前記グローバル共有メモリ部が該第 1 の相互結合網に接続され、

前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部と前記ローカル共有メモリ部の第 1 のメモリとが接続された第 2 の接続部と他の各ディスク制御クラスタ内の該第 2 の接続部が第 2 の相互結合網を介して接続されたことを特徴とするストレージシステム。

【請求項 7】 ホストコンピュータとのインターフェースを有する 1 または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する 1 または複数のディスクインターフェース部と、

前記ディスク装置に対しリード／ライトされるデータを格納する第 1 のメモリと、前記チャンネルインターフェース部及び前記ディスクインターフェース部と前記第 1 のメモリとの間のデータ転送に関する制御情報及び前記ディスク装置の管理情報を格納する第 2 のメモリとを有するローカル共有メモリ部とを有し、前記チャンネルインターフェース部は前記ホストコンピュータからのデータのリード／ライト要求に対し、前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部内の前記第 1 のメモリとの間のデータ転送を実行し、前記ディスクインターフェース部は、前記ディスク装置と前記ローカル共有メモリ部内の前記第 1 のメモリとの間のデータ転送を実行することにより、データのリード／ライ

トを行う複数のディスク制御クラスタと、

前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであって、

前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部の第2のメモリに前記ディスク制御クラスタ内で直接接続され、

該各ディスク制御クラスタ内の該チャネルインターフェース部と前記ディスクインターフェース部との第1の接続部と他の各ディスク制御クラスタ内の該第1の接続部が第1の相互結合網を介して接続され、

前記グローバル共有メモリ部が該第1の相互結合網に接続され、

前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部の第1のメモリに前記ディスク制御クラスタ内で直接接続され、

該各ディスク制御クラスタ内の該ローカル共有メモリ部の第1のメモリと他の各ディスク制御クラスタ内の該ローカル共有メモリ部の第1のメモリが第2の相互結合網を介して接続されたことを特徴とするストレージシステム。

【請求項8】 請求項1～4、6～7のいずれかの請求項記載のストレージシステムにおいて、

前記ローカル共有メモリ部は該ローカル共有メモリが属する前記ディスク制御クラスタが管理する記憶領域を示す情報を格納しており、前記グローバル共有メモリ部は前記各ディスク制御クラスタが管理する記憶領域を示す情報を格納しており、

前記チャネルインターフェース部内のプロセッサは、前記ホストコンピュータから前記ディスク制御クラスタの前記チャネルインターフェース部にデータのリード/ライト要求があった場合、該ディスク制御クラスタ内の前記ローカル共有メモリ部にアクセスし、該ディスク制御クラスタが管理する記憶領域内に前記要求のデータが格納されているか否か判定し、格納されていない場合、前記グローバル共有メモリ部へアクセスし、前記要求データが格納されているディスク制御クラスタを調べる手段を有することを特徴とするストレージシステム。

【請求項9】 ホストコンピュータとのインターフェースを有する1または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報及び前記ディスク装置の管理情報を格納するローカル共有メモリ部と、該チャンネルインターフェース部と該ディスクインターフェース部と該ローカル共有メモリ部とが接続された第1の接続部を有し、

前記チャンネルインターフェース部は前記ホストコンピュータからのデータのリード／ライト要求に対し、前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は、前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することによりデータのリード／ライトを行うディスク制御クラスタであって、

該ディスク制御クラスタの第1の接続部は、前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部が接続された第2の接続部への接続パスを有することを特徴とするディスク制御クラスタ。

【請求項10】 ホストコンピュータとのインターフェースを有する1または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報及び前記ディスク装置の管理情報を格納するローカル共有メモリ部とを有し、

該ローカル共有メモリ部に該チャンネルインターフェース部と該ディスクインターフェース部が接続され、前記チャンネルインターフェース部は前記ホストコンピュータからのデータのリード／ライト要求に対し、前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は、前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することにより、データのリード／ライトを行うディスク制御クラスタであって、

該ディスク制御クラスタのローカル共有メモリ部は、前記各ディスク制御クラ

スタの管理情報を格納するグローバル共有メモリ部が接続された第2の接続部への接続パスを有することを特徴とするディスク制御クラスタ。

【請求項11】 ホストコンピュータとのインターフェースを有する1または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報及び前記ディスク装置の管理情報を格納するローカル共有メモリ部と、該チャンネルインターフェース部と該ディスクインターフェース部とが接続された第1の接続部を有し、

該ローカル共有メモリ部に該チャンネルインターフェース部と該ディスクインターフェース部が接続され、前記チャンネルインターフェース部は前記ホストコンピュータからのデータのリード／ライト要求に対し、前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は、前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することによりデータのリード／ライトを行う複数のディスク制御クラスタと、

前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を備え、

該グローバル共有メモリ部は第2の接続部に接続パスで接続され、前記各ディスク制御クラスタの第1の接続部は該第2の接続部に接続パスで接続されたことを特徴とするストレージシステム。

【請求項12】 ホストコンピュータとのインターフェースを有する1または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、前記ディスク装置に対しリード／ライトされるデータと前記データの転送に関する制御情報及び前記ディスク装置の管理情報を格納するローカル共有メモリ部と、該チャンネルインターフェース部と該ディスクインターフェース部とが接続された第1の接続部を有し、

該ローカル共有メモリ部に該チャンネルインターフェース部と該ディスクインターフェース部が接続され、前記チャンネルインターフェース部は前記ホストコンピュータからのデータのリード／ライト要求に対し、前記ホストコンピュータとの

インターフェースと前記ローカル共有メモリ部との間のデータ転送を実行し、前記ディスクインターフェース部は、前記ディスク装置と前記ローカル共有メモリ部との間のデータ転送を実行することによりデータのリード／ライトを行う複数のディスク制御クラスタと、

前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を備え、

該グローバル共有メモリ部は第 2 の接続部に接続バスで接続され、前記各ディスク制御クラスタの第 1 の接続部は該第 2 の接続部に接続バスで接続されたストレージシステムにおいて、

前記各ディスク制御クラスタを実装した各ディスク制御クラスタの筐体に前記第 1 の接続部を接続した 1 以上の第 1 のコネクタを設け、

前記グローバル共有メモリ部と該グローバル共有メモリ部を接続した前記第 2 の接続部を実装した筐体に第 2 の接続部を接続した複数の第 2 のコネクタを設け、ストレージシステムを構成する前記各ディスク制御クラスタの第 1 のコネクタと前記第 2 のコネクタを接続バスにより接続し、

ストレージシステムにディスク制御クラスタを増設する時、増設したディスク制御クラスタの第 1 のコネクタを接続バスにより前記第 2 のコネクタに接続することを特徴とするディスク制御クラスタの増設方法。

【請求項 1 3】 請求項 1 2 記載のディスク制御クラスタの増設方法において、

前記グローバル共有メモリ部は、前記第 2 の接続部が有する前記各第 2 のコネクタに前記ディスク制御クラスタが接続されているか否かを示す第 1 のテーブルと、該第 2 のコネクタに接続されている前記ディスク制御クラスタが管理する記憶領域を示す第 2 のテーブルを格納しており、前記ディスク制御クラスタを増設するのに伴い、前記第 1、第 2 のテーブルに増設された前記ディスク制御クラスタに関する有効な情報を追加することを特徴とするディスク制御クラスタの増設方法。

【発明の詳細な説明】

【 0 0 0 1 】

【産業上の利用分野】

本発明は、データを複数のディスク装置に格納するストレージシステムとそれを構成するディスク制御クラスタに関する。

【0002】

【従来の技術】

半導体記憶装置を記憶媒体とするコンピュータの主記憶のI/O性能に比べて、磁気ディスクを記憶媒体とするディスクサブシステム（以下「サブシステム」という。）のI/O性能は3～4桁程度小さく、従来からこの差を縮めること、すなわちサブシステムのI/O性能を向上させる努力がなされている。

サブシステムのI/O性能を向上させるための1つの方法として、複数のディスク装置でサブシステムを構成し、データを複数のディスク装置に格納する、いわゆるディスクと呼ばれるシステムが知られている。

例えば、従来技術では、図2に示すようにホストコンピュータ3とディスク制御装置4との間のデータ転送を実行する複数のチャンネルIF部11と、ディスク装置2とディスク制御装置4間のデータ転送を実行する複数のディスクIF部16と、ディスク装置2のデータとディスク制御装置4に関する制御情報（例えば、ディスク制御装置4内のデータ転送制御に関する情報、ディスク装置2に格納するデータの管理情報）を格納する共有メモリ部20とを備え、1つのディスク制御装置4内において、共有メモリ部20は全てのチャンネルIF部11及びディスクIF部16からアクセス可能な構成となっている。

このディスク制御装置4では、チャンネルIF部11及びディスクIF部16と共有メモリ部20との間は相互結合網30で接続される。

チャンネルIF部11は、ホストコンピュータ3と接続するためのインターフェース及びホストコンピュータ3に対する入出力を制御するマイクロプロセッサ（図示せず）を有している。

また、ディスクIF部16は、ディスク装置2と接続するためのインターフェース及びディスク装置2に対する入出力を制御するマイクロプロセッサ（図示せず）を有している。また、ディスクIF部16は、RAID機能の実行も行う。

【0003】

インターネットの普及等により企業で扱うデータは爆発的に増大しており、データセンタ等では一台のディスク制御装置で扱えるデータ量以上のデータを記憶する必要がある。

このため、図2に示すようにディスク制御装置4を複数台設置し、それらのホストコンピュータ3とのインターフェースをSANスイッチ5を介して、ホストコンピュータ3に接続していた。

また、データ量の増大に伴いSANスイッチ5に接続するディスク制御装置4の台数が増えると、ホストコンピュータ3とSANスイッチ5を含めたシステム全体（このシステムをストレージ・エリア・ネットワーク（SAN）と呼ぶ）の管理が複雑化する。

それに対処するため、SANスイッチ5にSANアプライアンス6を接続し、SANアプライアンス6においてSANスイッチ5に繋がる全てのディスク制御装置4が管理するデータのディレクトリサービスを行い、ホストコンピュータ3に対して複数のディスク制御装置4を1つのストレージシステムに見せる処理、言い換えると、個々のディスク制御装置4が提供する記憶領域を1つの大きな記憶領域の固まりに見せ、その中から必要な量の記憶領域をホストコンピュータ3に割当てるという処理を行っていた。

【0004】

【発明が解決しようとする課題】

銀行、証券、電話会社等に代表される大企業では、従来各所に分散していたコンピュータ及びストレージを、データセンタの中に集中化してコンピュータシステム及びストレージシステム構成することにより、コンピュータシステム及びストレージシステムの運用、保守、管理に要する費用を削減する傾向にある。

このような傾向の中で、大型／ハイエンドのディスク制御装置には、数百台以上のホストコンピュータへ接続するためのチャネルインターフェースのサポート（コネクティビティ）、数百テラバイト以上の記憶容量のサポートが要求されている。

一方、近年のオープン市場の拡大、ストレージ・エリア・ネットワーク（SAN）の普及により、大型／ハイエンドのディスク制御装置と同様の高機能・高信頼性

を備えた小規模構成(小型筐体)のディスク制御装置への要求が高まっている。

前者の要求に対しては、従来の大型／ハイエンドのディスク制御装置を複数接続して超大規模なストレージシステムを構成する方法が考えられる。

また後者の要求に対しては、従来の大型／ハイエンドのディスク制御装置の最小構成のモデルにおいて筐体を小型化した装置を構成する方法が考えられる。

また、この小型化した装置を複数台接続することにより、従来のディスク制御装置がサポートしている中規模から大規模の構成をサポートするストレージシステムを構成する方法が考えられる。

ストレージシステムでは、上記のように、小規模な構成から超大規模な構成まで、同一の高機能・高信頼なアーキテクチャで対応可能な、スケーラビリティのある構成のシステムが必要となっており、そのためには、複数のディスク制御装置をクラスタリングし、1つのシステムとして運用できるストレージシステムが必要となる。

図2に示す従来技術では、複数のディスク制御装置4をSANスイッチ5を介してホストコンピュータ3に接続し、SANアプライアンス6によりホストコンピュータ3に対して複数のディスク制御装置4を1つのストレージシステムに見せていた。

しかし、SANアプライアンス6上で動作するソフトウェアで複数のディスク制御装置4を1つのシステムとして運用するため、従来の単体の大型のディスク制御装置に比べて信頼性、可用性が低いという問題があった。

また、SANアプライアンス6上でホストコンピュータ3から要求されたデータが存在するディスク制御装置4を検索するため、性能が低下するという問題があった。

本発明の目的は、小規模な構成から超大規模な構成まで、同一の高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティのある構成のストレージシステムを提供することにある。

より具体的には、本発明の目的は、複数台のディスク制御装置をまとめて1つのシステムとしたストレージシステムにおいて高信頼・高性能なシステムを提供することにある。

【 0 0 0 5 】

【課題を解決するための手段】

上記目的を達成するため、本発明は、ホストコンピュータとのインターフェースを有する1または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、ディスク装置に対しリード／ライトされるデータとデータの転送に関する制御情報及びディスク装置の管理情報を格納するローカル共有メモリ部とを有し、チャンネルインターフェース部はホストコンピュータからのデータのリード／ライト要求に対し、ホストコンピュータとのインターフェースとローカル共有メモリ部との間のデータ転送を実行し、ディスクインターフェース部は、ディスク装置とローカル共有メモリ部との間のデータ転送を実行することにより、データのリード／ライトを行う複数のディスク制御クラスタと、前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであり、

該各ディスク制御クラスタと他の各ディスク制御クラスタは相互結合網に接続され、該相互結合網に前記グローバル共有メモリ部が接続されたことによって達成される。

また、前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部と前記ローカル共有メモリ部とが接続された接続部と他の各ディスク制御クラスタ内の該接続部が前記相互結合網を介して接続されたことによって達成される。

また、前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部に前記ディスク制御クラスタ内で直接接続され、該各ディスク制御クラスタ内の該ローカル共有メモリ部と他の各ディスク制御クラスタ内の該ローカル共有メモリ部が前記相互結合網を介して接続されたことによって達成される。

また、前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部に前記ディスク制御クラスタ内で直接接続され、該各ディスク制御クラスタ内の該チャンネルインターフェース部と前記ディスクインターフェース部との接続部と他の各ディスク制御

クラスタ内の該接続部が前記相互結合網を介して接続されたことによって達成される。

また、ホストコンピュータとのインターフェースを有する1または複数のチャネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、該1または複数のチャネルインターフェース部と該1または複数のディスクインターフェース部を接続する接続部を有する複数のディスク制御クラスタと、ディスク装置に対しリード/ライトされるデータとデータの転送に関する制御情報、ディスク装置の管理情報、及びディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであり、

前記各ディスク制御クラスタ内の接続部と他の各ディスク制御クラスタ内の該接続部が相互結合網を介して接続され、該相互結合網に前記グローバル共有メモリ部が接続されたことによって達成される。

また、ホストコンピュータとのインターフェースを有する1または複数のチャネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、ディスク装置に対しリード/ライトされるデータを格納する第1のメモリと、チャネルインターフェース部及びディスクインターフェース部と第1のメモリとの間のデータ転送に関する制御情報及びディスク装置の管理情報を格納する第2のメモリとを有するローカル共有メモリ部とを有し、チャネルインターフェース部はホストコンピュータからのデータのリード/ライト要求に対し、前記ホストコンピュータとのインターフェースと前記ローカル共有メモリ部内の第1のメモリとの間のデータ転送を実行し、ディスクインターフェース部は、ディスク装置とローカル共有メモリ部内の第1のメモリとの間のデータ転送を実行することにより、データのリード/ライトを行う複数のディスク制御クラスタと、前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであり、

前記各ディスク制御クラスタ内の前記チャネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部の第2のメモリに前記ディスク制御クラスタ内で直接接続され、

該各ディスク制御クラスタ内の該チャンネルインターフェース部と前記ディスクインターフェース部との第1の接続部と他の各ディスク制御クラスタ内の該第1の接続部が第1の相互結合網を介して接続され、

前記グローバル共有メモリ部が該第1の相互結合網に接続され、

前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部と前記ローカル共有メモリ部の第1のメモリとが接続された第2の接続部と他の各ディスク制御クラスタ内の該第2の接続部が第2の相互結合網を介して接続されたことによって達成される。

また、ホストコンピュータとのインターフェースを有する1または複数のチャンネルインターフェース部と、ディスク装置とのインターフェースを有する1または複数のディスクインターフェース部と、ディスク装置に対しリード/ライトされるデータを格納する第1のメモリと、チャンネルインターフェース部及びディスクインターフェース部と第1のメモリとの間のデータ転送に関する制御情報及び前記ディスク装置の管理情報を格納する第2のメモリとを有するローカル共有メモリ部とを有し、チャンネルインターフェース部は前記ホストコンピュータからのデータのリード/ライト要求に対し、ホストコンピュータとのインターフェースとローカル共有メモリ部内の第1のメモリとの間のデータ転送を実行し、ディスクインターフェース部は、ディスク装置とローカル共有メモリ部内の第1のメモリとの間のデータ転送を実行することにより、データのリード/ライトを行う複数のディスク制御クラスタと、前記各ディスク制御クラスタの管理情報を格納するグローバル共有メモリ部を有するストレージシステムであり、

前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディスクインターフェース部が前記ローカル共有メモリ部の第2のメモリに前記ディスク制御クラスタ内で直接接続され、

該各ディスク制御クラスタ内の該チャンネルインターフェース部と前記ディスクインターフェース部との第1の接続部と他の各ディスク制御クラスタ内の該第1の接続部が第1の相互結合網を介して接続され、

前記グローバル共有メモリ部が該第1の相互結合網に接続され、

前記各ディスク制御クラスタ内の前記チャンネルインターフェース部と前記ディ

スクインターフェース部が前記ローカル共有メモリ部の第1のメモリに前記ディスク制御クラスタ内で直接接続され、

該各ディスク制御クラスタ内の該ローカル共有メモリ部の第1のメモリと他の各ディスク制御クラスタ内の該ローカル共有メモリ部の第1のメモリが第2の相互結合網を介して接続されたことによって達成される。

【0006】

【発明の実施形態】

以下、本発明の実施例を図面を用いて説明する。

《実施例1》

図1、図3、図12、及び図13に、本発明の一実施例を示す。

以下の実施例において、相互結合網はスイッチを利用したものを例にして説明してあるが、相互に接続され制御情報やデータが転送されれば良いのであり、例えばバスで構成されても良い。

図1に示すように、ストレージシステム1は複数のディスク制御クラスタ1-1乃至1-nから構成される。

ディスク制御クラスタ1-1は、ホストコンピュータ3とのインターフェース部（チャンネルIF部）11と、ディスク装置2とのインターフェース部（ディスクIF部）16と、ローカル共有メモリ部22を有し、チャンネルIF部11及びディスクIF部16とローカル共有メモリ部22の間は複数のディスク制御クラスタ1-1乃至1-nに跨る相互結合網31を介して接続され、グローバル共有メモリ部21は相互結合網31に接続されている。

すなわち、相互結合網31を介して、全てのチャンネルIF部11及びディスクIF部12から、グローバル共有メモリ部22へアクセス可能な構成となっている。

【0007】

チャンネルIF部11の具体的な一例を図12に示す。

チャンネルIF部11は、ホストコンピュータ3との2つのIF（ホストIF）202と、ホストコンピュータ3に対する入出力を制御する2つのマイクロプロセッサ201と、グローバル共有メモリ部21あるいはローカル共有メモリ部2

2 へのアクセスを制御するアクセス制御部（メモリアクセス制御部）2 0 6 を有し、ホストコンピュータ 3 とグローバル共有メモリ部 2 1 あるいはローカル共有メモリ部 2 2 間のデータ転送、及びマイクロプロセッサ 2 0 1 とグローバル共有メモリ部 2 1 あるいはローカル共有メモリ部 2 2 間の制御情報の転送を実行する。

マイクロプロセッサ 2 0 1 及びホスト I F 2 0 2 は内部バス 2 0 5 によって接続され、メモリアクセス制御部 2 0 6 は 2 つのホスト I F 2 0 2 に直接接続され、また内部バス 2 0 5 に接続されている。

【 0 0 0 8 】

ディスク I F 部 1 6 の具体的な一例を図 1 3 に示す。

ディスク I F 部 1 6 は、ディスク装置 2 との 2 つの I F （ドライブ I F ） 2 0 3 と、ディスク装置 2 に対する入出力を制御する 2 つのマイクロプロセッサ 2 0 1 と、グローバル共有メモリ部 2 1 あるいはローカル共有メモリ部 2 2 へのアクセスを制御するアクセス制御部（メモリアクセス制御部） 2 0 6 を有し、ディスク装置 2 とグローバル共有メモリ部 2 1 あるいはローカル共有メモリ部 2 2 間のデータ転送、及びマイクロプロセッサ 2 0 1 とグローバル共有メモリ部 2 1 あるいはローカル共有メモリ部 2 2 間の制御情報の転送を実行する。

マイクロプロセッサ 2 0 1 及びドライブ I F 2 0 3 は内部バス 2 0 5 によって接続され、メモリアクセス制御部 2 0 6 は 2 つのドライブ I F 2 0 3 に直接接続され、また、内部バス 2 0 5 に接続されている。

ディスク I F 部 1 6 は RAID 機能の実行も行う。

1 つのディスク制御クラスタは 1 つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体 1 つのディスク制御装置として機能を備えているものである。

【 0 0 0 9 】

ストレージシステムの具体的な一例を図 3 に示す。

ストレージシステム 1 は、複数のディスク制御クラスタ 1 - 1 乃至 1 - n と、グローバル共有メモリ部 2 1 と、2 つのグローバルスイッチ（G S W ） 1 1 5 と、アクセスバス 1 3 6 と、アクセスバス 1 3 7 を有する。

グローバルスイッチ 1 1 5 は、グローバル共有メモリ部 2 1 からのバスと複数のディスク制御クラスタからのバスを接続する接続部である。

ディスク制御クラスタ 1 - 1 乃至 1 - n は、ホストコンピュータ 3 との 2 つのチャンネル I F 部 1 1 と、ディスク装置 2 との 2 つのディスク I F 部 1 6 と、2 つのローカルスイッチ (L S W) 1 1 0 と、2 つのローカル共有メモリ部 2 2 と、アクセスパス 1 3 1 と、アクセスパス 1 3 2 と、アクセスパス 1 3 6 を有する。

グローバル共有メモリ部 2 1 は、メモリコントローラ 1 0 1 とメモリモジュール 1 0 5 とを有し、ディスク制御クラスタ 1 - 1 乃至 1 - n の管理情報 (例えば、各ディスク制御クラスタが管理する記憶領域の情報や、ディスク制御クラスタの稼動状況及び構成情報等) を格納する。

ローカルスイッチ 1 1 0 は、チャンネル I F 部からのバスと、ディスク I F 部からのバスと、ローカル共有メモリからのバスを接続する接続部である。

ローカル共有メモリ部 2 2 は、メモリコントローラ 1 0 0 とメモリモジュール 1 0 5 とを有し、ディスク制御クラスタの制御情報 (例えば、チャンネル I F 部 1 1 及びディスク I F 部 1 6 とローカル共有メモリ部 2 2 との間のデータ転送制御に関する情報、ディスク装置 2 に記録するデータの管理情報等) とディスク装置 2 に記録するデータを格納する。

【 0 0 1 0 】

チャンネル I F 部 1 1 内のメモリアクセス制御部 2 0 6 には 2 本のアクセスパス 1 3 1 を接続し、それらを 2 つの異なる L S W 1 1 0 にそれぞれ接続する。

L S W 1 1 0 には 2 本のアクセスパス 1 3 2 を接続し、それらを 2 つの異なるローカル共有メモリ部 2 2 内のメモリコントローラ 1 0 0 にそれぞれ接続する。

したがって、メモリコントローラ 1 0 0 には、2 つの L S W 1 1 0 から 1 本ずつ、計 2 本のアクセスパス 1 3 2 が接続される。

こうすることにより、1 つのメモリアクセス制御部 2 0 6 から 1 つのメモリコントローラ 1 0 0 へのアクセスルートが 2 つとなる。

これにより、1 つのアクセスパスまたは L S W 1 1 0 に障害が発生した場合でも、もう 1 つのアクセスルートによりローカル共有メモリ部 2 2 へアクセスすることが可能となるため、耐障害性を向上させることができる。

【0011】

L SW110には、2つのチャンネルIF部11と、2つのディスクIF部16からそれぞれ1本ずつ、計4本のアクセスパス131が接続される。

また、L SW110には、2つのローカル共有メモリ部22へのアクセスパス132が2本とG SW115へのアクセスパス136が1本接続される。

L SW110には上記のようなアクセスパスが接続されるため、L SW110内では、チャンネルIF部11及びディスクIF部16からの4本のアクセスパスからの要求を、自ディスク制御クラスタ内のローカル共有メモリ部22への2本のアクセスパスと、G SW115への1本のアクセスパス136に振分ける機能を有する。

【0012】

G SW115には、各ディスク制御クラスタから1本ずつ、ディスク制御クラスタ数分の本数のアクセスパス136が接続される。

また、G SW115には、2つのグローバル共有メモリ部21内のメモリコントローラ101へのアクセスパス137が1本ずつ、計2本接続される。

こうすることにより、1つのメモリアクセス制御部206から1つのメモリコントローラ101へのアクセスルートが2つとなる。

これにより、1つのアクセスパスまたはL SW110またはG SW115に障害が発生した場合でも、もう1つのアクセスルートによりグローバル共有メモリ部21へアクセスすることが可能となるため、耐障害性を向上させることができる。

【0013】

G SW115を使わずに、アクセスパス136をメモリコントローラ101に直接接続しても本発明を実施する上で問題ない。

そうすることにより、G SW115で発生するデータ転送処理のオーバーヘッドを削減することが可能となり、性能が向上する。

G SW115を使わない場合、1つのメモリアクセス制御部206から1つのメモリコントローラ101へのアクセスルートを2つ確保し、耐障害性を向上するためには、L SW110にアクセスパス136を2本接続し、それぞれを異な

るメモリコントローラ 1 0 1 へ接続する。

【 0 0 1 4 】

図 3 で L S W 1 1 0 はチャンネル I F 部 1 1 及びディスク I F 部 1 6 とローカル共有メモリ部 2 2 との接続部であり、G S W 1 1 5 はディスク制御クラスタ 1 - 1 乃至 1 - n とグローバル共有メモリ部 2 1 との接続部である。

図 3 において、G S W 1 1 5 とグローバル共有メモリ部 2 1 をボックスに実装し、モジュール化した各ディスク制御クラスタ 1 - 1 乃至 1 - n といっしょに、1 つの筐体の中に実装しても良い。

また、各ディスク制御クラスタ 1 - 1 乃至 1 - n を別個の筐体として、距離的に離れた場所に分散しても良い。

【 0 0 1 5 】

図 3 において、ディスク制御装置 1 - 1 に接続されたホストコンピュータ 3 からストレージシステム 1 に記録されたデータを読み出す場合の一例を述べる。

まず、ホストコンピュータ 3 は、自身が接続されているディスク制御クラスタ 1 - 1 内のチャンネル I F 部 1 1 にデータの読出し要求を発行する。

要求を受けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、自ディスク制御クラスタ 1 - 1 内のローカル共有メモリ部 2 2 にアクセスし、要求されたデータがどのディスク装置 2 内に格納されているかを調べる。

ローカル共有メモリ部 2 2 には、要求データのアドレスとそのデータが実際に記録されているディスク装置 2 内のアドレスを対応させる変換テーブルが格納されており、要求されたデータがどのディスク装置 2 内に格納されているかを調べることができる。

【 0 0 1 6 】

要求されたデータが自ディスク制御クラスタ 1 - 1 に接続されたディスク装置 2 に格納されていた場合、要求を受けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、さらに自ディスク制御クラスタ 1 - 1 内のローカル共有メモリ部 2 2 にアクセスし、要求されたデータがローカル共有メモリ部 2 2 内に格納されているかどうかを確認する。

ローカル共有メモリ部 2 2 にはディスク装置 2 に格納するデータとともにその

データのディレクトリ情報が格納されており、ローカル共有メモリ部 2 2 内に要求データが存在するかどうかを確認できる。

それにより自ディスク制御クラスタ 1 - 1 のローカル共有メモリ部 2 2 内にデータがあった場合は、ローカル共有メモリ部 2 2 にアクセスしてそのデータを自身の L S W 1 1 0 を介してチャンネル I F 部 1 1 まで転送し、さらにホストコンピュータ 3 に送る。

【 0 0 1 7 】

自ディスク制御クラスタ 1 - 1 のローカル共有メモリ部 2 2 内にデータが存在しなかった場合は、チャンネル I F 部 1 1 のマイクロプロセッサ 2 0 1 は、要求データが格納されているディスク装置 2 が接続されているディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 に対し、要求データを読み出しローカル共有メモリ部 2 2 に格納するというデータ要求の処理内容を示す制御情報を発行し、この制御情報の発行を受けたディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 は、要求データが格納されているディスク装置 2 からデータを読み出し、L S W 1 1 0 を介して、自ディスク制御クラスタ 1 - 1 内のローカル共有メモリ部 2 2 に要求データを転送し格納する。

すなわち、チャンネル I F 部 1 1 のマイクロプロセッサ 2 0 1 は、上記データ要求の処理内容を示す制御情報を発行し、ローカル共有メモリ部 2 2 の制御情報領域（ジョブ制御ブロック）に格納する。

ディスク I F 部 1 6 のマイクロプロセッサ 2 0 1 は、ローカル共有メモリ部 2 2 の制御情報領域をポーリングで監視し、上記発行された制御情報が上記制御情報領域（ジョブ制御ブロック）に存在した場合は、要求データが格納されているディスク装置 2 からデータを読み出し、L S W 1 1 0 を介して、自ディスク制御クラスタ 1 - 1 内のローカル共有メモリ部 2 2 に要求データを転送し格納する。

【 0 0 1 8 】

ディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 は、要求データをローカル共有メモリ部 2 2 へ格納した後、前記制御情報を発行したチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 にローカル共有メモリ部 2 2 内のデータを格納したアドレスを、ローカル共有メモリ部 2 2 内の制御情報を介して伝える。それを受

けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、ローカル共有メモリ部 2 2 からデータを読み出し、ホストコンピュータ 3 に送る。

すなわち、ディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 は、要求データをローカル共有メモリ部 2 2 へ格納した後、処理の実行の終了とデータを格納したアドレスを示す制御情報を発行し、ローカル共有メモリ部 2 2 の制御情報領域に格納する。

前記制御情報を発行したチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、ローカル共有メモリ部 2 2 の制御情報領域をポーリングで監視し、ディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 から発行された制御情報が存在する場合、ローカル共有メモリ部内のデータを格納したアドレスによりローカル共有メモリ部 2 2 からデータを読み出し、チャンネル I F 部 1 1 まで転送し、さらにホストコンピュータ 3 に送る。

【0019】

要求されたデータが自ディスク制御クラスタ 1-1 に接続されたディスク装置 2 に格納されていなかった場合、要求を受けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、グローバル共有メモリ部 2 1 にアクセスし、要求されたデータが格納されているディスク装置が接続されているディスク制御クラスタを調べる。

グローバル共有メモリ部 2 2 には、要求データのアドレスとそのデータが格納されているディスク装置が接続されたディスク制御クラスタを対応させる変換テーブルが格納されており、要求されたデータがどのディスク制御クラスタに格納されているかを調べることができる。

【0020】

要求されたデータがディスク制御クラスタ 1-n に接続されたディスク装置 2 に格納されていた場合、要求を受けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、グローバル共有メモリ部 2 1 を介してディスク制御クラスタ 1-n に要求データをディスク制御クラスタ 1-n 内のローカル共有メモリ部 2 2 へ格納するように要求する。

グローバル共有メモリ部 2 1 の中には、ディスク制御クラスタ間でデータ要求

の受け渡しを行う制御情報を格納する領域があり、その領域は要求先のディスク制御クラスタ毎に分割されており、要求を受けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、ディスク制御クラスタ 1 - n の制御情報を格納する領域に要求データをディスク制御クラスタ 1 - n 内のローカル共有メモリ部 2 2 へ格納するよう要求する情報を要求データのアドレスとともに格納する。

ディスク制御クラスタ 1 - n 内のチャンネル I F 部 1 1 またはディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 は、グローバル共有メモリ 2 1 内の自ディスク制御クラスタへの制御情報を格納する領域をポーリングで監視している。

【 0 0 2 1 】

データを要求する制御情報の領域にデータ要求があった場合、ディスク制御クラスタ 1 - n 内のチャンネル I F 部 1 1 またはディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 は、要求されたデータが自ディスク制御クラスタ 1 - n 内のローカル共有メモリ部 2 2 内に格納されているかどうかを確認する。

ローカル共有メモリ部 2 2 にはディスク装置 2 に格納するデータとともにそのデータのディレクトリ情報が格納されており、ローカル共有メモリ部 2 2 内に要求データが存在するかどうかを確認できる。

それにより自ディスク制御クラスタ 1 - n のローカル共有メモリ部 2 2 内にデータがあった場合は、上記と同様にして、グローバル共有メモリ 2 1 を介してディスク制御クラスタ 1 - 1 内のデータ要求を受けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 にディスク制御クラスタ 1 - n 内のローカル共有メモリ部 2 2 へ格納したことを伝える。

それを受け、ディスク制御クラスタ 1 - 1 内のデータ要求を受けたチャンネル I F 部 1 1 内のマイクロプロセッサ 2 0 1 は、G S W 1 1 5 及び L S W 1 1 0 を介してディスク制御クラスタ 1 - n のローカル共有メモリ部 2 2 から要求データを読み出し、チャンネル I F 部 1 1 まで転送し、さらにホストコンピュータ 3 に送る。

【 0 0 2 2 】

ディスク制御クラスタ 1 - n のローカル共有メモリ部 2 2 内にデータが存在しなかった場合は、ディスク制御クラスタ 1 - n のチャンネル I F 部 1 1 またはディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 は、要求データが格納されている

ディスク装置 2 が接続されているディスク I F 部 1 6 内のマイクロプロセッサ 2 0 1 に対し、要求データを読み出し、ローカル共有メモリ部 2 2 に格納するデータ要求の処理内容を示す制御情報を発行し、ローカル共有メモリ部 2 2 の制御情報領域（ジョブ制御ブロック）に格納する。

ディスク I F 部 1 6 のマイクロプロセッサ 2 0 1 は、ローカル共有メモリ部 2 2 の制御情報領域をポーリングで監視し、上記発行された制御情報が上記制御情報領域（ジョブ制御ブロック）に存在した場合は、要求データが格納されているディスク装置 2 からデータを読み出し、L S W 1 1 0 を介して、自ディスク制御クラスタ 1 - n 内のローカル共有メモリ部 2 2 に要求データを転送し格納する。

その後の処理は、上記ローカル共有メモリ部 2 2 内に要求データがあった場合の処理と同様である。

【 0 0 2 3 】

本実施例によれば、ホストコンピュータ 3 は要求データがどのディスク制御クラスタに繋がるディスク装置 2 に格納されているかを意識することなく、自身が繋がるディスク制御クラスタにアクセス要求を発行するだけで、データの書き込み及び読み出しを行うことが可能になり、ホストコンピュータ 3 に対して、複数台のディスク制御クラスタ 1 - 1 乃至 1 - n を 1 つのストレージシステムに見せることが可能となる。

そして、ディスク制御クラスタが 1 個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティのある構成のストレージシステムを提供することが可能となる。

【 0 0 2 4 】

《 実施例 2 》

図 4、図 5、図 1 2、及び図 1 3 に、本発明の一実施例を示す。

図 4 に示すように、ディスク制御ユニット 1 - 1 乃至 1 - n からなるストレージシステム 1 の構成は、チャンネル I F 部 1 1 及びディスク I F 部 1 6 とローカル共有メモリ部 2 2 及び相互結合網 3 1 の間の接続構成を除いて、実施例 1 の図 1 に示す構成と同様である。

チャンネル I F 部 1 1 及びディスク I F 部 1 6 とローカル共有メモリ部 2 2 の間は、ディスク制御クラスタ内では直接接続されている。

また、複数のディスク制御クラスタ 1 - 1 乃至 1 - n 間では、ローカル共有メモリ部 2 2 は相互結合網 3 1 を介して接続されており、その相互結合網 3 1 にグローバル共有メモリ 2 1 が接続されている。

【 0 0 2 5 】

上記のように、この実施例ではディスク制御ユニット 1 - 1 乃至 1 - n においてチャンネル I F 部 1 1 及びディスク I F 部 1 6 とローカル共有メモリ部 2 2 を直接接続することにより、実施例 1 で示した相互結合網 3 1 を介して接続する場合に比べ、ローカル共有メモリ部 2 2 へのアクセス時間を短縮することが可能になる。

チャンネル I F 部 1 1 及びディスク I F 部 1 6 の構成は、それぞれ図 1 2、図 1 3 に示す実施例 1 の構成と同様である。

1 つのディスク制御クラスタは 1 つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体 1 つのディスク制御装置として機能を備えているものである。

【 0 0 2 6 】

ストレージシステム 1 の具体的な一例を図 5 に示す。

ディスク制御クラスタ 1 - 1 乃至 1 - n 内の構成も、チャンネル I F 部 1 1 及びディスク I F 部 1 6 とローカル共有メモリ部 2 2 の間の接続構成とディスク制御クラスタ 1 - 1 乃至 1 - n と G S W 1 1 5 の接続構成を除いて、実施例 1 の図 3 に示す構成と同様である。

ストレージシステム 1 は、複数のディスク制御クラスタ 1 - 1 乃至 1 - n と、グローバル共有メモリ部 2 1 と、2 つのグローバルスイッチ (G S W) 1 1 5 と、アクセスパス 1 3 6 と、アクセスパス 1 3 7 を有する。

ディスク制御クラスタ 1 - 1 乃至 1 - n は、ホストコンピュータ 3 との 2 つのチャンネル I F 部 1 1 と、ディスク装置 2 との 2 つのディスク I F 部 1 6 と、2 つのローカル共有メモリ部 2 2 と、アクセスパス 1 3 3 と、アクセスパス 1 3 6 を有する。

【0027】

チャンネルIF部11内のメモリアクセス制御部206には2本のアクセスパス133を接続し、それらを2つの異なるメモリコントローラ100にそれぞれ接続する。

したがって、メモリコントローラ100には、2つのチャンネルIF部11と2つのディスクIF部16から1本ずつ、計4本のアクセスパス133が接続される。また、GSW115へのアクセスパス136が1本接続される。

メモリコントローラ100には上記のようなアクセスパスが接続されるため、メモリコントローラ100内では、チャンネルIF部11及びディスクIF部16からの4本のアクセスパス133からの要求を、メモリモジュール105への1本のアクセスパスと、GSW115への1本のアクセスパス136に振分ける機能を有する。

【0028】

実施例1と同様にGSW115を使わずに、アクセスパス136をメモリコントローラ101に直接接続しても本発明を実施する上で問題ない。そうすることにより、GSW115で発生するデータ転送処理のオーバーヘッドを削減することが可能となり、性能が向上する。

GSW115を使わない場合、1つのメモリコントローラ100から1つのメモリコントローラ101へのアクセスルートを2つ確保し、耐障害性を向上するためには、メモリコントローラ100にアクセスパス136を2本接続し、それぞれを異なるメモリコントローラ101へ接続する。

また、実施例1と同様に、図5において、GSW115とグローバル共有メモリ部21をボックスに実装し、モジュール化した各ディスク制御クラスタ1-1乃至1-nといっしょに、1つの筐体の中に実装しても良い。また、各ディスク制御クラスタ1-1乃至1-nを別個の筐体として、距離的に離れた場所に分散しても良い。

【0029】

本実施例において、ホストコンピュータ3からストレージシステムへのデータの読み出し／書き込みを行う場合の、ストレージシステム1内の各部の動作は、

チャンネル I F 部 1 1 及びディスク I F 部 1 6 からローカル共有メモリ部 2 2 へのアクセスが直接になることと、チャンネル I F 部 1 1 及びディスク I F 部 1 6 からグローバル共有メモリ部 2 1 へのアクセスがメモリコントローラ 1 0 0 を介して行われることを除いて、実施例 1 と同様である。

本実施例によれば、ホストコンピュータ 3 は要求データがどのディスク制御クラスタに繋がるディスク装置 2 に格納されているかを意識することなく、自身が繋がるディスク制御クラスタにアクセス要求を発行するだけで、データの書き込み及び読出しを行うことが可能になり、ホストコンピュータ 3 に対して、複数台のディスク制御クラスタ 1 - 1 乃至 1 - n を 1 つのストレージシステムに見せることが可能となる。

そして、ディスク制御クラスタが 1 個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティのある構成のストレージシステムを提供することが可能となる。

【 0 0 3 0 】

《実施例 3》

図 6、図 7、図 1 2、及び図 1 3 に、本発明の一実施例を示す。

図 6 に示すように、ディスク制御ユニット 1 - 1 乃至 1 - n からなるストレージシステム 1 の構成は、チャンネル I F 部 1 2 及びディスク I F 部 1 7 とローカル共有メモリ部 2 2 の間の接続構成を除いて、実施例 1 の図 1 に示す構成と同様である。

チャンネル I F 部 1 2 及びディスク I F 部 1 7 とローカル共有メモリ部 2 2 の間は、ディスク制御クラスタ内では直接接続されている。

また、複数のディスク制御クラスタ 1 - 1 乃至 1 - n 間では、チャンネル I F 部 1 2 及びディスク I F 部 1 7 が相互結合網 3 1 を介して接続されており、その相互結合網 3 1 にグローバル共有メモリ 2 1 が接続されている。

上記のように、この実施例ではディスク制御ユニット 1 - 1 乃至 1 - n においてチャンネル I F 部 1 2 及びディスク I F 部 1 7 とローカル共有メモリ部 2 2 を直接接続することにより、実施例 1 で示した相互結合網 3 1 を介して接続する場

合に比べ、ローカル共有メモリ部22へのアクセス時間を短縮することが可能になる。

チャンネルIF部12及びディスクIF部17の構成は、それぞれ図12、図13に示すチャンネルIF部11及びディスクIF部16の構成において、メモリアクセス制御部206のアクセスパスを4本に増やした構成となる。

ここで、4本のアクセスパスの内、2本はアクセスパス131、もう2本がアクセスパス133（図7参照）となる。

1つのディスク制御クラスタは1つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体1つのディスク制御装置として機能を備えているものである。

【0031】

ストレージシステム1の具体的な一例を図7に示す。

ディスク制御クラスタ1-1乃至1-n内の構成も、チャンネルIF部12及びディスクIF部17とローカル共有メモリ部22の間の接続構成を除いて、実施例1の図3に示す構成と同様である。

ストレージシステム1は、複数のディスク制御クラスタ1-1乃至1-nと、グローバル共有メモリ部21と、2つのグローバルスイッチ（GSW）115と、アクセスパス136と、アクセスパス137を有する。

ディスク制御クラスタ1-1乃至1-nは、ホストコンピュータ3との2つのチャンネルIF部12と、ディスク装置2との2つのディスクIF部17と、2つのローカルスイッチ（LSW）110と、2つのローカル共有メモリ部22と、アクセスパス131と、アクセスパス133と、アクセスパス136を有する。

ローカルスイッチ110は、チャンネルIF部からのパスと、ディスクIF部からのパスを接続する接続部である。

チャンネルIF部12及びディスクIF部17内のメモリアクセス制御部206には2本のアクセスパス133を接続し、それらを2つの異なるメモリコントローラ100にそれぞれ接続する。したがって、メモリコントローラ100には、2つのチャンネルIF部11と2つのディスクIF部16から1本ずつ、計4本のアクセスパス133が接続される。

さらに、チャンネル I F 部 1 2 及びディスク I F 部 1 7 内のメモリアクセス制御部 2 0 6 には 2 本のアクセスパス 1 3 1 を接続し、それらを 2 つの異なる L S W 1 1 0 にそれぞれ接続する。したがって、L S W 1 1 0 には、2 つのチャンネル I F 部 1 2 と 2 つのディスク I F 部 1 7 から 1 本ずつ、計 4 本のアクセスパス 1 3 1 が接続される。また、G S W 1 1 5 へのアクセスパス 1 3 6 が 1 本接続される。

【0032】

実施例 1 と同様に G S W 1 1 5 を使わずに、アクセスパス 1 3 6 をメモリコントローラ 1 0 1 に直接接続しても本発明を実施する上で問題ない。そうすることにより、G S W 1 1 5 で発生するデータ転送処理のオーバーヘッドを削減することが可能となり、性能が向上する。G S W 1 1 5 を使わない場合、1 つの L S W 1 1 0 から 1 つのメモリコントローラ 1 0 1 へのアクセスルートを 2 つ確保し、耐障害性を向上するためには、L S W 1 1 0 にアクセスパス 1 3 6 を 2 本接続し、それぞれを異なるメモリコントローラ 1 0 1 へ接続する。

また、実施例 1 と同様に、図 7 において、G S W 1 1 5 とグローバル共有メモリ部 2 1 をボックスに実装し、モジュール化した各ディスク制御クラスタ 1 - 1 乃至 1 - n といっしょに、1 つの筐体の中に実装しても良い。また、各ディスク制御クラスタ 1 - 1 乃至 1 - n を別個の筐体として、距離的に離れた場所に分散しても良い。

【0033】

本実施例において、ホストコンピュータ 3 からストレージシステムへのデータの読み出し／書き込みを行う場合の、ストレージシステム 1 内の各部の動作は、チャンネル I F 部 1 2 及びディスク I F 部 1 7 からローカル共有メモリ部 2 2 へのアクセスが直接になることを除いて、実施例 1 と同様である。

【0034】

本実施例によれば、ホストコンピュータ 3 は要求データがどのディスク制御クラスタに繋がるディスク装置 2 に格納されているかを意識することなく、自身が繋がるディスク制御クラスタにアクセス要求を発行するだけで、データの書き込み及び読出しを行うことが可能になり、ホストコンピュータ 3 に対して、複数台

のディスク制御クラスタ 1-1 乃至 1-n を 1 つのストレージシステム 1 に見せることが可能となる。そして、ディスク制御クラスタが 1 個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティのある構成のストレージシステムを提供することが可能となる。

【 0 0 3 5 】

《実施例 4》

図 8、図 9、図 12、及び図 13 に、本発明の一実施例を示す。

図 8 に示すように、ディスク制御ユニット 1-1 乃至 1-n からなるストレージシステム 1 の構成は、実施例 1 においてローカル共有メモリ 22 を除いた構成である。

このため、実施例 1 の各ディスク制御ユニット 1-1 乃至 1-n のローカル共有メモリ 22 に格納する情報を全てグローバル共有メモリ 21 に格納する。

複数のディスク制御クラスタ 1-1 乃至 1-n 間では、チャンネル I/F 部 11 及びディスク I/F 部 16 が相互結合網 31 を介して接続されており、その相互結合網 31 にグローバル共有メモリ 21 が接続されている。

チャンネル I/F 部 11 及びディスク I/F 部 16 の構成は、それぞれ図 12、図 13 に示す実施例 1 の構成と同様である。

1 つのディスク制御クラスタは 1 つの筐体として構成されるか、またはモジュールとして構成されても良い。

【 0 0 3 6 】

ストレージシステム 1 の具体的な一例を図 9 に示す。

ディスク制御クラスタ 1-1 乃至 1-n 内の構成も、ローカル共有メモリ部 22 が無いことを除いて、実施例 1 の図 3 に示す構成と同様である。

ストレージシステム 1 は、複数のディスク制御クラスタ 1-1 乃至 1-n と、グローバル共有メモリ部 21 と、2 つのグローバルスイッチ (G SW) 115 と、アクセスパス 136 と、アクセスパス 137 を有する。

ディスク制御クラスタ 1-1 乃至 1-n は、ホストコンピュータ 3 との 2 つのチャンネル I/F 部 11 と、ディスク装置 2 との 2 つのディスク I/F 部 16 と、2 つ

のローカルスイッチ (LSW) 110と、アクセスパス131と、アクセスパス136を有する。

ローカルスイッチ110は、チャンネルIF部からのパスと、ディスクIF部からのパスを接続する接続部である。

チャンネルIF部11及びディスクIF部16内のメモリアクセス制御部206には2本のアクセスパス131を接続し、それらを2つの異なるLSW110にそれぞれ接続する。

したがって、LSW110には、2つのチャンネルIF部11と2つのディスクIF部16から1本ずつ、計4本のアクセスパス131が接続される。また、GSW115へのアクセスパス136が1本接続される。

【0037】

実施例1と同様にGSW115を使わずに、アクセスパス136をメモリコントローラ101に直接接続しても本発明を実施する上で問題ない。そうすることにより、GSW115で発生するデータ転送処理のオーバーヘッドを削減することが可能となり、性能が向上する。

GSW115を使わない場合、1つのLSW110から1つのメモリコントローラ101へのアクセスルートを2つ確保し、耐障害性を向上するためには、LSW110にアクセスパス136を2本接続し、それぞれを異なるメモリコントローラ101へ接続する。

また実施例1と同様に、図9において、GSW115とグローバル共有メモリ部21をボックスに実装し、モジュール化した各ディスク制御クラスタ1-1乃至1-nといっしょに、1つの筐体の中に実装しても良い。また、各ディスク制御クラスタ1-1乃至1-nを別個の筐体として、距離的に離れた場所に分散しても良い。

【0038】

本実施例において、ホストコンピュータ3からストレージシステムへのデータの読み出し／書き込みを行う場合の、ストレージシステム1内の各部の動作は、実施例1の処理においてローカル共有メモリ22における処理を全てグローバル共有メモリ21で行うことを除いて、実施例1と同様である。

【 0 0 3 9 】

本実施例によれば、ホストコンピュータ 3 は要求データがどのディスク制御クラスタに繋がるディスク装置 2 に格納されているかを意識することなく、自身が繋がるディスク制御クラスタにアクセス要求を発行するだけで、データの書き込み及び読出しを行うことが可能になり、ホストコンピュータ 3 に対して、複数台のディスク制御クラスタ 1-1 乃至 1-n を 1 つのストレージシステム 1 に見せることが可能となる。

そして、ディスク制御クラスタが 1 個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティのある構成のストレージシステムを提供することが可能となる。

【 0 0 4 0 】

《実施例 5》

図 10 に、本発明の一実施例を示す。

以下の実施例において、相互結合網はスイッチを利用したものを例にして説明してあるが、相互に接続され制御情報やデータが転送されれば良いのであり、例えばバスで構成されても良い。

図 10 に示すように、ストレージシステム 1 は複数のディスク制御クラスタ 1-1 乃至 1-n から構成される。

ディスク制御クラスタ 1-1 乃至 1-n は、ホストコンピュータ 3 とのインターフェース部（チャンネル I F 部） 13 と、ディスク装置 2 とのインターフェース部（ディスク I F 部） 18 と、メモリ 1 : 25 とメモリ 2 : 26 を有するローカル共有メモリ部 22 を有し、チャンネル I F 部 13 及びディスク I F 部 18 とメモリ 2 の間は、ディスク制御クラスタ内部では直接接続される。

また、チャンネル I F 部 13 及びディスク I F 部 18 は複数のディスク制御クラスタ 1-1 乃至 1-n に跨る相互結合網 1 : 32 を介して接続され、グローバル共有メモリ部 21 は相互結合網 1 : 32 に接続される。すなわち、相互結合網 1 : 32 を介して、全てのチャンネル I F 部 13 及びディスク I F 部 18 から、グローバル共有メモリ部 21 へアクセス可能な構成となっている。

また、チャンネル I F 部 1 3 及びディスク I F 部 1 8 とメモリ 1 の間は、複数のディスク制御クラスタ 1 - 1 乃至 1 - n に跨る相互結合網 2 : 3 3 を介して接続される。

【 0 0 4 1 】

チャンネル I F 部 1 3 の具体的な一例を図 1 4 に示す。

チャンネル I F 部 1 3 は、ホストコンピュータ 3 との 2 つの I F (ホスト I F) 2 0 2 と、ホストコンピュータ 3 に対する入出力を制御する 2 つのマイクロプロセッサ 2 0 1 と、グローバル共有メモリ部 2 1 あるいはメモリ 2 : 2 6 へのアクセスを制御するアクセス制御部 1 (メモリアクセス制御部 1) 2 0 7 と、メモリ 1 : 2 5 へのアクセスを制御するアクセス制御部 2 (メモリアクセス制御部 2) 2 0 8 とを有し、ホストコンピュータ 3 とメモリ 1 間のデータ転送、及びマイクロプロセッサ 2 0 1 とグローバル共有メモリ部 2 1 あるいはメモリ 2 間の制御情報の転送を実行する。

マイクロプロセッサ 2 0 1 及びホスト I F 2 0 2 は内部バス 2 0 5 によって接続され、メモリアクセス制御部 2 0 7 は内部バス 2 0 5 に接続され、メモリアクセス制御部 2 0 8 は 2 つのホスト I F 2 0 2 に直接接続され、また内部バス 2 0 5 に接続されている。

【 0 0 4 2 】

ディスク I F 部 1 8 の具体的な一例を図 1 5 に示す。

ディスク I F 部 1 8 は、ディスク装置 2 との 2 つの I F (ドライブ I F) 2 0 3 と、ディスク装置 2 に対する入出力を制御する 2 つのマイクロプロセッサ 2 0 1 と、グローバル共有メモリ部 2 1 あるいはメモリ 2 へのアクセスを制御するアクセス制御部 1 (メモリアクセス制御部 1) 2 0 7 と、メモリ 1 へのアクセスを制御するアクセス制御部 2 (メモリアクセス制御部 2) 2 0 8 とを有し、ディスク装置 2 とメモリ 1 間のデータ転送、及びマイクロプロセッサ 2 0 1 とグローバル共有メモリ部 2 1 あるいはメモリ 2 間の制御情報の転送を実行する。

マイクロプロセッサ 2 0 1 及びドライブ I F 2 0 3 は内部バス 2 0 5 によって接続され、メモリアクセス制御部 2 0 7 は内部バス 2 0 5 に接続され、メモリアクセス制御部 2 0 8 は 2 つのドライブ I F 2 0 3 に直接接続され、また内部バス

205に接続されている。ディスクIF部18はRAID機能の実行も行う。

1つのディスク制御クラスタは1つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体1つのディスク制御装置として機能を備えているものである。

【0043】

ストレージシステムの具体的な一例は、チャンネルIF部13及びディスクIF部18とメモリ2:26と相互結合網1:32とグローバル共有メモリ21との接続構成は、実施例3の図7に示す構成と同様になる。

また、チャンネルIF部13及びディスクIF部18とメモリ1と相互結合網2:33との接続構成は、実施例1の図3に示す構成においてグローバル共有メモリ21を除いた構成と同様になる。

【0044】

グローバル共有メモリ部21は、ディスク制御クラスタ1-1乃至1-nの管理情報（例えば、各ディスク制御クラスタが管理する記憶領域の情報や、ディスク制御クラスタの稼動状況及び構成情報等）を格納する。

メモリ1は、ディスク装置2に記録するデータを一時的に格納する。

また、メモリ2は、ディスク制御クラスタの制御情報（例えば、チャンネルIF部13及びディスクIF部18とメモリ1:25との間のデータ転送制御に関する情報、ディスク装置2に記録するデータの管理情報等）を格納する。

図10において、相互結合網1:32を形成するディスク制御クラスタ外のスイッチ及び相互結合網2:33を形成するディスク制御クラスタ外のスイッチとグローバル共有メモリ部21をボックスに実装し、モジュール化した各ディスク制御クラスタ1-1乃至1-nと一緒に、1つの筐体の中に実装しても良い。また、各ディスク制御クラスタ1-1乃至1-nを別個の筐体として、距離的に離れた場所に分散しても良い。

【0045】

図10において、ディスク制御装置1-1に接続されたホストコンピュータ3からストレージシステム1に記録されたデータを読み出す場合の一例を述べる。

まず、ホストコンピュータ3は、自身が接続されているディスク制御クラスタ

1-1内のチャンネルIF部13にデータの読出し要求を発行する。

要求を受けたチャンネルIF部13内のマイクロプロセッサ201は、自ディスク制御クラスタ1-1内のメモリ2:26にアクセスし、要求されたデータがどのディスク装置2内に格納されているかを調べる。メモリ2:26には、要求データのアドレスとそのデータが実際に記録されているディスク装置2内のアドレスを対応させる変換テーブルが格納されており、要求されたデータがどのディスク装置2内に格納されているかを調べることができる。

要求されたデータが自ディスク制御クラスタ1-1に接続されたディスク装置2に格納されていた場合、要求を受けたチャンネルIF部13内のマイクロプロセッサ201は、さらに自ディスク制御クラスタ1-1内のメモリ2:26にアクセスし、要求されたデータがメモリ1:25に格納されているかどうかを確認する。メモリ2:26にはメモリ1:25に格納されているデータのディレクトリ情報が格納されており、メモリ1:25に要求データが存在するかどうかを確認できる。

それにより自ディスク制御クラスタ1-1のメモリ1:25にデータがあった場合は、そのデータをチャンネルIF部13まで転送し、ホストコンピュータ3に送る。

【0046】

自ディスク制御クラスタ1-1のメモリ1:25にデータが存在しなかった場合は、チャンネルIF部13内のマイクロプロセッサ201は要求データが格納されているディスク装置2が接続されているディスクIF部18内のマイクロプロセッサ201に対し、要求データを読出し、メモリ1:25に格納するように、メモリ2:26の制御情報を発行し、この制御情報の発行を受けたディスクIF部18内のマイクロプロセッサ201は、要求データが格納されているディスク装置2からデータを読出し、自ディスク制御クラスタ1-1内のメモリ1:25に要求データを転送し格納する。

すなわち、チャンネルIF部13のマイクロプロセッサ201は、上記データ要求の処理内容を示す制御情報を発行し、メモリ部2:26の制御情報領域（ジョブ制御ブロック）に格納する。

ディスク I F 部 18 のマイクロプロセッサ 201 は、メモリ部 2 : 26 の制御情報領域をポーリングで監視し、上記発行された制御情報が上記制御情報領域（ジョブ制御ブロック）に存在した場合は、要求データが格納されているディスク装置 2 からデータを読み出し、自ディスク制御クラスタ 1-1 内のメモリ 1 : 25 に要求データを転送し格納する。

ディスク I F 部 18 内のマイクロプロセッサ 201 は、要求データをメモリ 1 : 25 へ格納した後、制御情報を発行したチャンネル I F 部 13 内のマイクロプロセッサ 201 に、メモリ 1 : 25 内のデータを格納したアドレスを、メモリ 2 : 26 内の制御情報を介して伝える。それを受けたチャンネル I F 部 13 内のマイクロプロセッサ 201 は、メモリ 1 : 25 からデータを読み出し、ホストコンピュータ 3 へ送る。

すなわち、ディスク I F 部 18 内のマイクロプロセッサ 201 は、要求データをメモリ 1 : 25 へ格納した後、処理の実行の終了とデータを格納したアドレスを示す制御情報を発行し、メモリ 2 : 26 の制御情報領域に格納する。

前記制御情報を発行したチャンネル I F 部 13 内のマイクロプロセッサ 201 は、メモリ 2 : 26 の制御情報領域をポーリングで監視し、ディスク I F 部 18 内のマイクロプロセッサ 201 から発行された制御情報が存在する場合、メモリ 1 : 25 内のデータを格納したアドレスによりメモリ 1 : 25 からデータを読み出し、チャンネル I F 部 13 まで転送し、さらにホストコンピュータ 3 に送る。

【0047】

要求されたデータが自ディスク制御クラスタ 1-1 に接続されたディスク装置 2 に格納されていなかった場合、要求を受けたチャンネル I F 部 13 内のマイクロプロセッサ 201 は、相互結合網 1 : 32 を介してグローバル共有メモリ部 21 にアクセスし、要求されたデータが格納されているディスク装置 2 が接続されているディスク制御クラスタを調べる。

グローバル共有メモリ部 21 には、要求データのアドレスとそのデータが格納されているディスク装置が接続されたディスク制御クラスタを対応させる変換テーブルが格納されており、要求されたデータがどのディスク制御クラスタに格納されているかを調べることができる。

【0048】

要求されたデータがディスク制御クラスタ1-nに接続されたディスク装置2に格納されていた場合、要求を受けたチャンネルIF部13内のマイクロプロセッサ201は、グローバル共有メモリ部21を介してディスク制御クラスタ1-nに要求データをディスク制御クラスタ1-n内のメモリ1:25へ格納するように要求する。

グローバル共有メモリ部21の中には、ディスク制御クラスタ間でデータ要求の受け渡しを行う制御情報を格納する領域があり、その領域は要求先のディスク制御クラスタ毎に分割されており、要求を受けたチャンネルIF部13内のマイクロプロセッサ201は、ディスク制御クラスタ1-nの領域に要求データをディスク制御クラスタ1-n内のメモリ1:25へ格納するよう要求する情報を要求データのアドレスとともに格納する。

ディスク制御クラスタ1-n内のディスクIF部18内のマイクロプロセッサ201は、グローバル共有メモリ21内の自ディスク制御クラスタへの要求領域をポーリングで監視している。

データを要求する制御情報を格納する要求領域にデータ要求があった場合、要求されたデータが自ディスク制御クラスタ1-n内のメモリ1:25内に格納されているかどうかを確認する。

メモリ2:26内にはメモリ1:25に格納されたデータのディレクトリ情報が格納されており、メモリ1:25内に要求データが存在するかどうかを確認できる。

それにより自ディスク制御クラスタ1-nのメモリ1:25内にデータがあった場合は、上記と同様にして、グローバル共有メモリ21を介してディスク制御クラスタ1-1内のデータ要求を受けたチャンネルIF部13内のマイクロプロセッサ201にディスク制御クラスタ1-n内のメモリ1:25へ格納したことを伝える。

それを受け、ディスク制御クラスタ1-1内のデータ要求を受けたチャンネルIF部13内のマイクロプロセッサ201は、相互結合網2:33を介してディスク制御クラスタ1-n内のメモリ1:25から要求データを転送し、ホストコン

ピュータ3に送る。

【0049】

ディスク制御クラスタ1-nのメモリ1:25内にデータが存在しなかった場合は、ディスク制御クラスタ1-nのディスクIF部18内のマイクロプロセッサ201は、要求データが格納されているディスク装置2が接続されているディスクIF部18内のマイクロプロセッサ201に対し、要求データを読み出し、メモリ1:25に格納するデータ要求の処理内容を示す制御情報を発行し、メモリ2:26の制御情報領域（ジョブ制御ブロック）に格納する。

ディスクIF部18内のマイクロプロセッサ201は、メモリ2:26の制御情報領域をポーリングで監視し、上記発行された制御情報が上記制御情報領域（ジョブ制御ブロック）に存在した場合は、要求データが格納されているディスク装置2からデータを読み出し、相互結合網2:33を介して、自ディスク制御クラスタ1-n内のメモリ1:25に要求データを転送し格納する。

その後の処理は、上記メモリ1:25に要求データがあった場合の処理と同様である。

【0050】

制御情報とデータはデータ長が数千倍異なるため、1回のデータ転送時間かなり異なる。このため、同じ相互結合網及びメモリを用いた場合、両者が互いの転送を妨げる。本実施例によれば、制御情報を転送する相互結合網1:32とデータを転送する相互結合網2:33を分けることができるため、両者が互いの転送を妨げることがなくなるため、性能が向上する。

【0051】

《実施例6》

図11に、本発明の一実施例を示す。

図11に示すように、ディスク制御ユニット1-1乃至1-nからなるストレージシステム1の構成は、チャンネルIF部13及びディスクIF部18とメモリ1:25及び相互結合網2:33の間の接続構成を除いて、実施例5の図10に示す構成と同様である。

チャンネルIF部13及びディスクIF部18とメモリ1:25の間は、ディス

ク制御クラスタ内では直接接続されている。また、複数のディスク制御クラスタ 1-1 乃至 1-n 間では、メモリ 1:25 は相互結合網 2:33 を介して接続される。

【0052】

上記のように、この実施例ではディスク制御ユニット 1-1 乃至 1-n 内においてチャンネル I/F 部 13 及びディスク I/F 部 18 とメモリ 1:25 を直接接続することにより、実施例 5 で示した相互結合網 2:33 を介して接続する場合に比べ、メモリ 1:25 へのアクセス時間を短縮することが可能になる。

チャンネル I/F 部 13 及びディスク I/F 部 18 の構成は、それぞれ図 14、図 15 に示す実施例 5 の構成と同様である。

1 つのディスク制御クラスタは 1 つの筐体として構成されるか、またはモジュールとして構成されても良いが、それ自体 1 つのディスク制御装置として機能を備えているものである。

【0053】

ストレージシステム 1 の具体的な一例は、チャンネル I/F 部 13 及びディスク I/F 部 18 とメモリ 2:26 と相互結合網 1:32 とグローバル共有メモリ 21 との接続構成は、実施例 3 の図 7 に示す構成と同様になる。また、チャンネル I/F 部 13 及びディスク I/F 部 18 とメモリ 1:25 と相互結合網 2:33 との接続構成は、実施例 2 の図 5 に示す構成においてグローバル共有メモリ部 21 を除いた構成と同様になる。

グローバル共有メモリ部 21 は、ディスク制御クラスタ 1-1 乃至 1-n の管理情報（例えば、各ディスク制御クラスタが管理する記憶領域の情報や、ディスク制御クラスタの稼動状況及び構成情報等）を格納する。

メモリ 1:25 は、ディスク装置 2 に記録するデータを一時的に格納する。

また、メモリ 2:26 は、ディスク制御クラスタの制御情報（例えば、チャンネル I/F 部 13 及びディスク I/F 部 18 とメモリ 1:25 との間のデータ転送制御に関する情報、ディスク装置 2 に記録するデータの管理情報等）を格納する。

【0054】

図 11 において、相互結合網 1:32 を形成するディスク制御クラスタ外のス

イッチ及び相互結合網 2 : 3 3 を形成するディスク制御クラスタ外のスイッチとグローバル共有メモリ部 2 1 をボックスに実装し、モジュール化した各ディスク制御クラスタ 1 - 1 乃至 1 - n といっしよに、1 つの筐体の中に実装しても良い。また、各ディスク制御クラスタ 1 - 1 乃至 1 - n を別個の筐体として、距離的に離れた場所に分散しても良い。

本実施例において、ホストコンピュータ 3 からストレージシステムへのデータの読み出し／書き込みを行う場合の、ストレージシステム 1 内の各部の動作は、チャンネル I F 部 1 3 及びディスク I F 部 1 8 からメモリ 1 : 2 5 へのアクセスが直接になることと、チャンネル I F 部 1 3 及びディスク I F 部 1 8 から他のディスク制御クラスタのメモリ 1 へのアクセスがメモリ 1 : 2 5 のメモリコントローラ(図示せず)を介して行われることを除いて、実施例 5 と同様である。

制御情報とデータはデータ長が数千倍異なるため、1 回のデータ転送時間がかなり異なる。このため、同じ相互結合網及びメモリを用いた場合、両者が互いの転送を妨げる。本実施例によれば、制御情報を転送する相互結合網 1 : 3 2 とデータを転送する相互結合網 2 : 3 3 を分けることができるため、両者が互いの転送を妨げることがなくなるため、性能が向上する。

【 0 0 5 5 】

《実施例 7》

図 1 6 ~ 図 1 8 に、実施例 1 のストレージシステム 1 におけるディスク制御クラスタの増設手順の一例を示す。

図 1 6 に示すように、スイッチボックス 3 1 0 は、別筐体として実装されている。

スイッチボックス 3 1 0 内には、G S W 1 1 5 とグローバル共有メモリ 2 1 が実装されている。

スイッチボックス 3 1 0 はコネクタ 3 2 1、コネクタ 3 2 2 をそれぞれ 8 個有し、ディスク制御クラスタを 8 クラスタ接続することができる。図ではディスク制御クラスタを 3 クラスタ接続した場合について示している。

G S W 1 1 5 のアクセスパス 1 3 6 は 1 本ずつコネクタ 3 2 1、コネクタ 3 2 2 に接続される。上記個数は一実施例に過ぎず、個数を上記に限定するものでは

ない。

各ディスク制御クラスタ1-1乃至1-3は、それぞれ筐体301乃至303に実装されている。筐体301乃至303はコネクタ321、コネクタ322を有し、2本のアクセスパス136がそれぞれに1本ずつ接続されている。

スイッチボックス310に、ケーブル331、ケーブル332を介してそれぞれのコネクタ321、コネクタ322により筐体301、302、303が接続される。

【0056】

ストレージシステム1において、ディスク制御クラスタを増設する場合は、次の手順による。

スイッチボックス310にディスク制御クラスタを増設するコネクタに余分があれば、そのコネクタにケーブル331、ケーブル332を接続する。

余分がなければ、GSWのみを実装したスイッチボックスを用意し、スイッチボックスを多段に接続した上でそのコネクタにケーブル331、ケーブル332を接続する。

それと共に、図17に示すGSW115のポートに接続されるディスク制御クラスタを示す、言い換えるとストレージシステム1を構成しているディスク制御クラスタを示すGSWポートークラスタ対応テーブル400と、図18に示すディスク制御クラスタが管理する論理ボリュームを示すGSWポートークラスタ対応テーブル405とを書き換える。

GSWポートークラスタ対応テーブル400とGSWポートークラスタ対応テーブル405はグローバル共有メモリ部21に格納されており、サービスプロセッサ(SVP)により書き換えることが可能である。

SVPは通常ノートパソコンであることが多く、ノートパソコンのディスプレイ上に図17及び図18に示すテーブルが表示され、そこで内容を書き換える。

【0057】

図17及び図18は、それぞれディスク制御クラスタの増設前、増設後のGSWポートークラスタ対応テーブル400及びGSWポートークラスタ対応テーブル405を示している。

ここでは、ストレージシステム 1 が増設前に 5 台のディスク制御クラスタで構成されており、そこに 1 台のディスク制御クラスタを増設する例を示している。

図 1 7 に示すように、G S W ポート番号 4 0 1 の 4 番ポートが未接続となっており、そのポートにクラスタ 5 のケーブルを接続した後、S V P のディスプレイ上でポート番号 4 の行のクラスタ番号 4 0 2 の列の未接続表示を 5 に書き換える。

その後、図 1 8 に示すように、クラスタ番号 4 0 2 のクラスタ 5 の行の論理ボリューム番号 4 0 6 の列の未接続表示を 1 6 6 4 0 ~ 2 0 7 3 5 に書き換える。ここで、論理ボリューム番号 4 0 6 は各クラスタが管理する論理ボリュームの範囲を示している。

増設前の論理ボリューム番号の最大値は 1 6 6 3 9 で、ディスク制御クラスタは 4 0 9 6 個の論理ボリュームを持っているため、ディスク制御クラスタ 5 の管理する論理ボリュームの範囲は 1 6 6 4 0 ~ 2 0 7 3 5 となる。論理ボリューム番号は連続せず飛び飛びになっても問題ない。

上記のようにすることで、ストレージシステムに新たにディスク制御クラスタを増設することができる。

【 0 0 5 8 】

【発明の効果】

本発明によれば、複数台のディスク制御クラスタを 1 つのシステムとして運用するストレージシステムにおいて、ディスク制御クラスタが 1 個だけの小規模な構成からディスク制御クラスタが数十個接続された超大規模な構成まで、ディスク制御クラスタ単体が持つ高信頼・高性能なアーキテクチャで対応可能な、スケラビリティのある構成のストレージシステムを提供することが可能となる。

【図面の簡単な説明】

【図 1】

本発明によるストレージシステムの実施例 1 の構成を示す図である。

【図 2】

従来の複数のディスク制御装置の構成を示す図である。

【図 3】

図 1 に示す実施例 1 のストレージシステムの詳細構成を示す図である。

【図 4】

本発明によるストレージシステムの実施例 2 の構成を示す図である。

【図 5】

図 4 に示す実施例 2 のストレージシステムの詳細構成を示す図である。

【図 6】

本発明によるストレージシステムの実施例 3 の構成を示す図である。

【図 7】

図 6 に示す実施例 3 のストレージシステムの詳細構成を示す図である。

【図 8】

本発明によるストレージシステムの実施例 4 の構成を示す図である。

【図 9】

図 8 に示す実施例 4 のストレージシステムの詳細構成を示す図である。

【図 1 0】

本発明によるストレージシステムの実施例 5 の構成を示す図である。

【図 1 1】

本発明によるストレージシステムの実施例 6 の構成を示す図である。

【図 1 2】

本発明によるストレージシステムを構成するチャネルインターフェース部の構成を示す図である。

【図 1 3】

本発明によるストレージシステムを構成するディスクインターフェース部の構成を示す図である。

【図 1 4】

本発明によるストレージシステムを構成するチャネルインターフェース部の他の構成を示す図である。

【図 1 5】

本発明によるストレージシステムを構成するディスクインターフェース部の他の構成を示す図である。

【図 16】

本発明によるディスク制御クラスタの増設方法を説明するための図である。

【図 17】

グローバル共有メモリ部内に格納されたストレージシステムの構成情報の一例を示す図である。

【図 18】

グローバル共有メモリ部内に格納されたストレージシステムの構成情報の他の一例を示す図である。

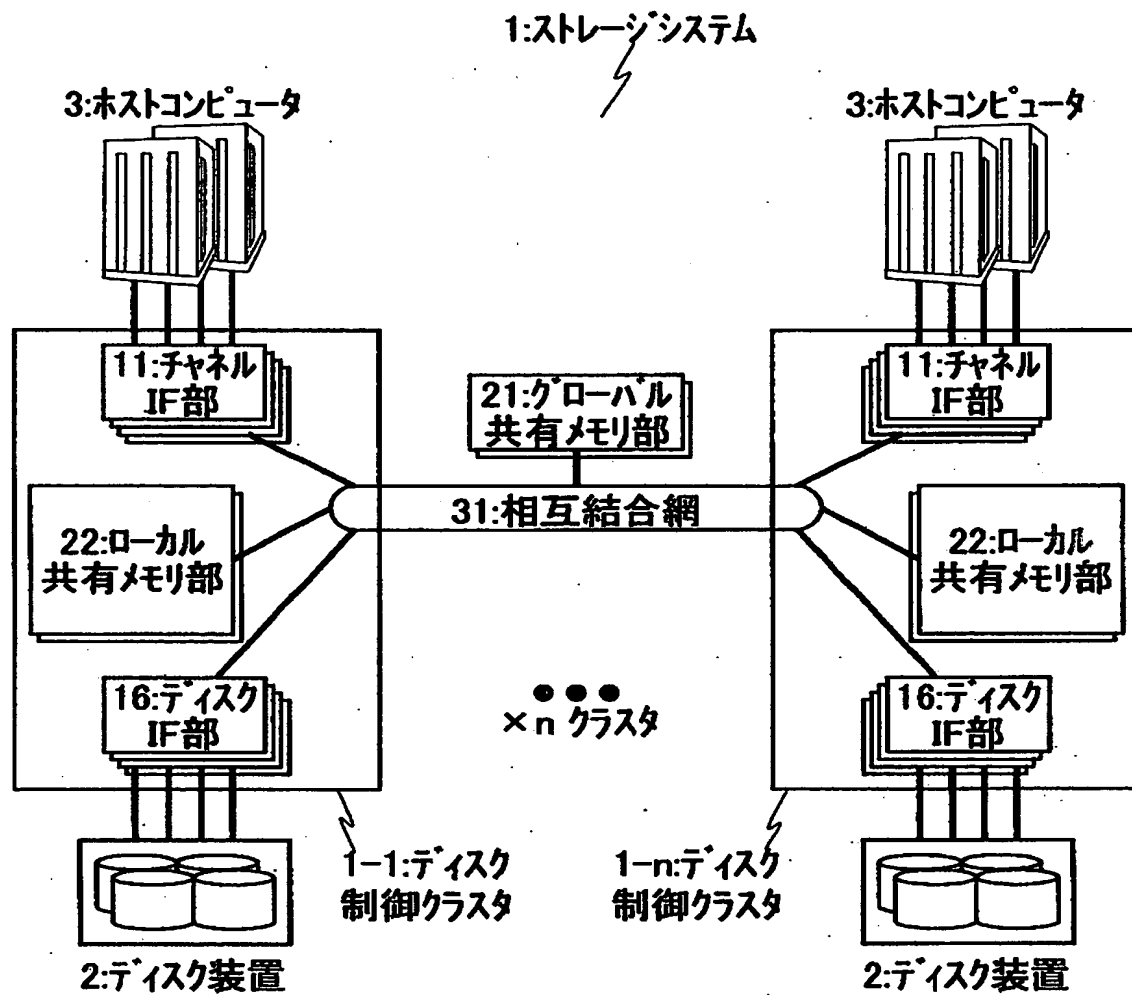
【符号の説明】

- 1 ストレージシステム
- 1-1、1-n ディスク制御クラスタ
- 2 ディスク装置
- 3 ホストコンピュータ
- 11、12、13 チャネル I/F 部
- 16、17、18 ディスク I/F 部
- 21 グローバル共有メモリ部
- 22 ローカル共有メモリ部
- 31、32、33 相互結合網

【書類名】 図面

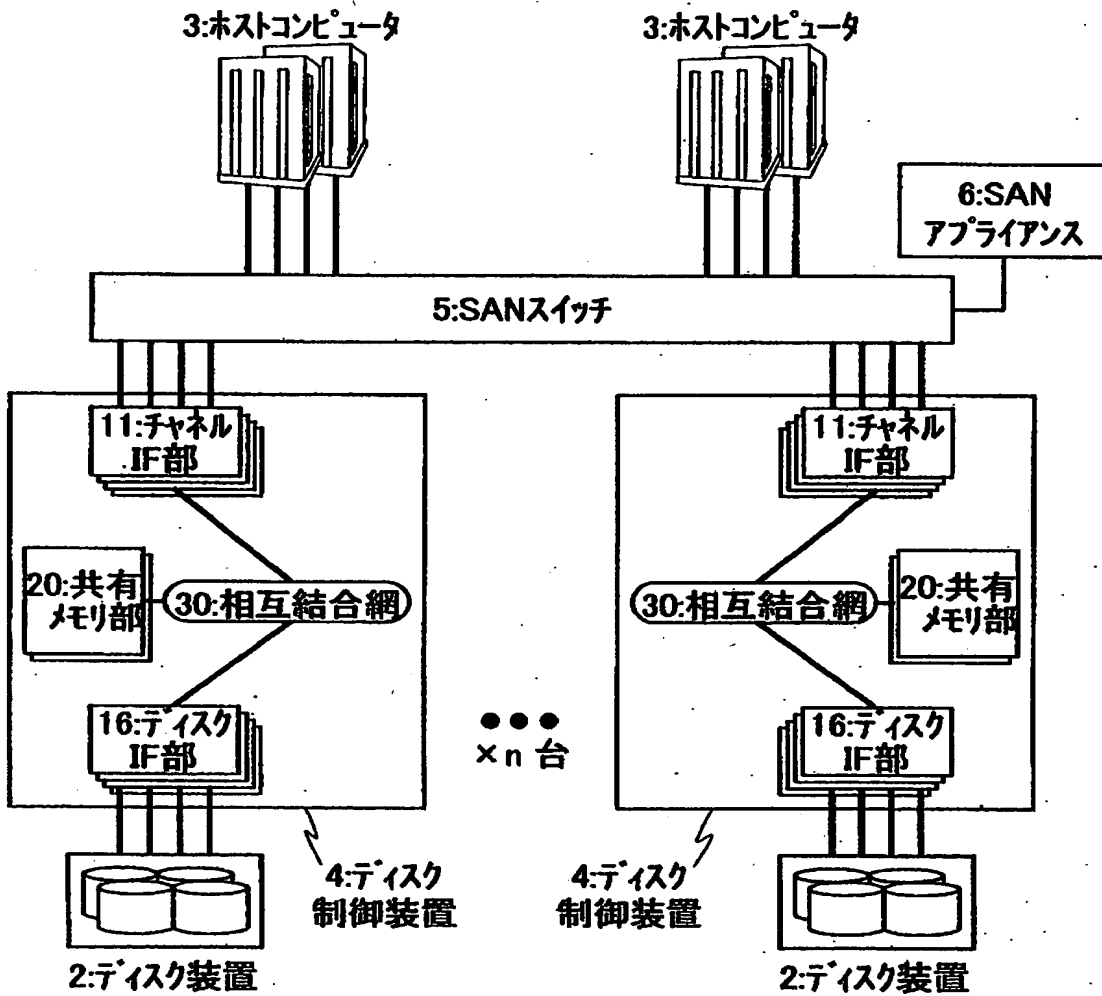
【図1】

図1

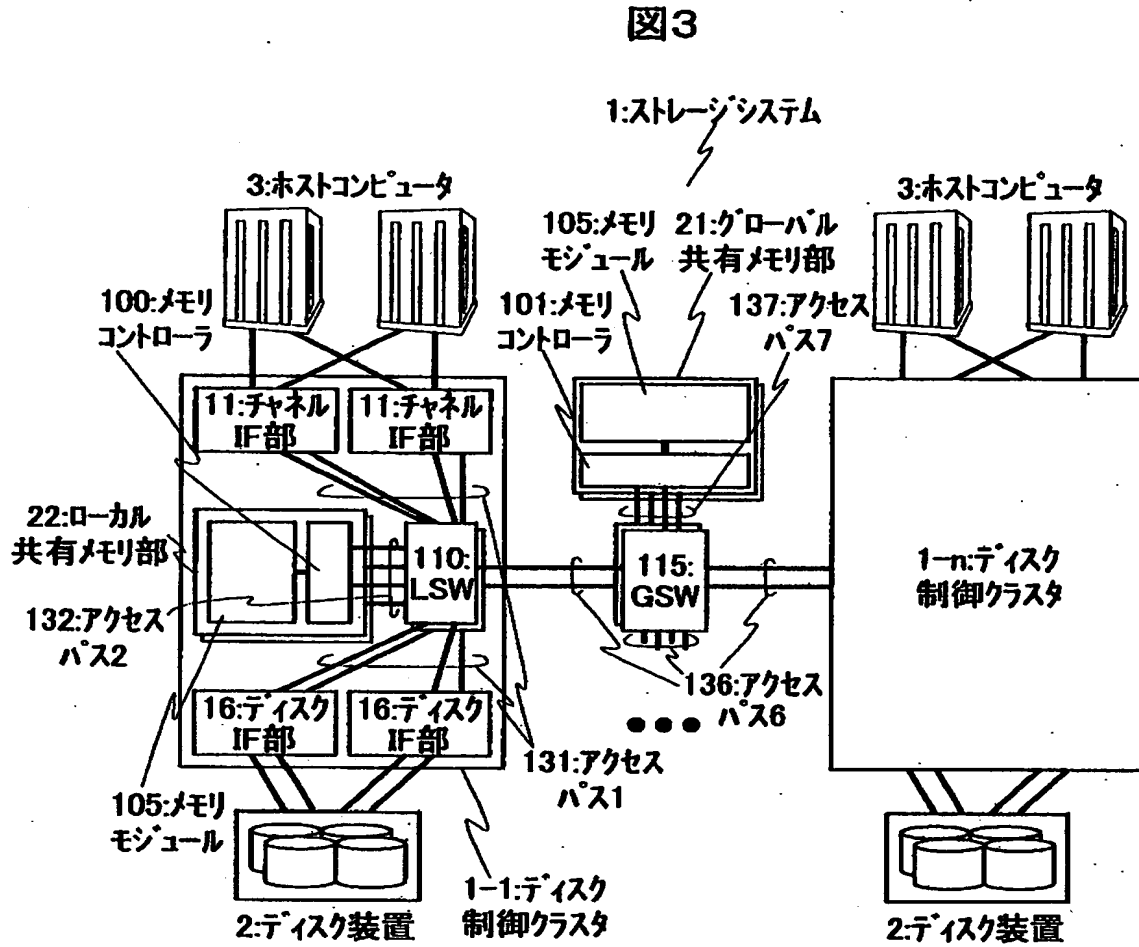


【図2】

図2

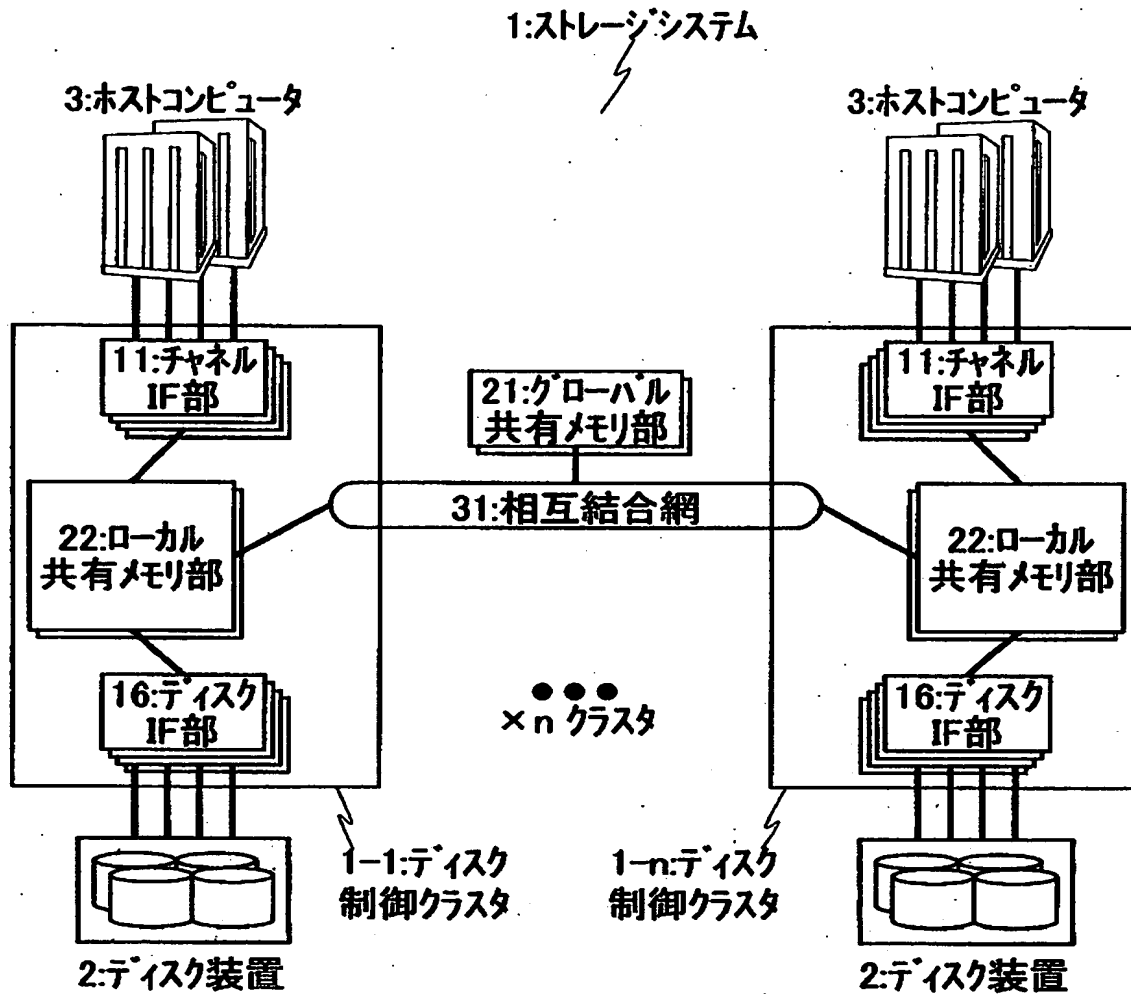


【図 3】

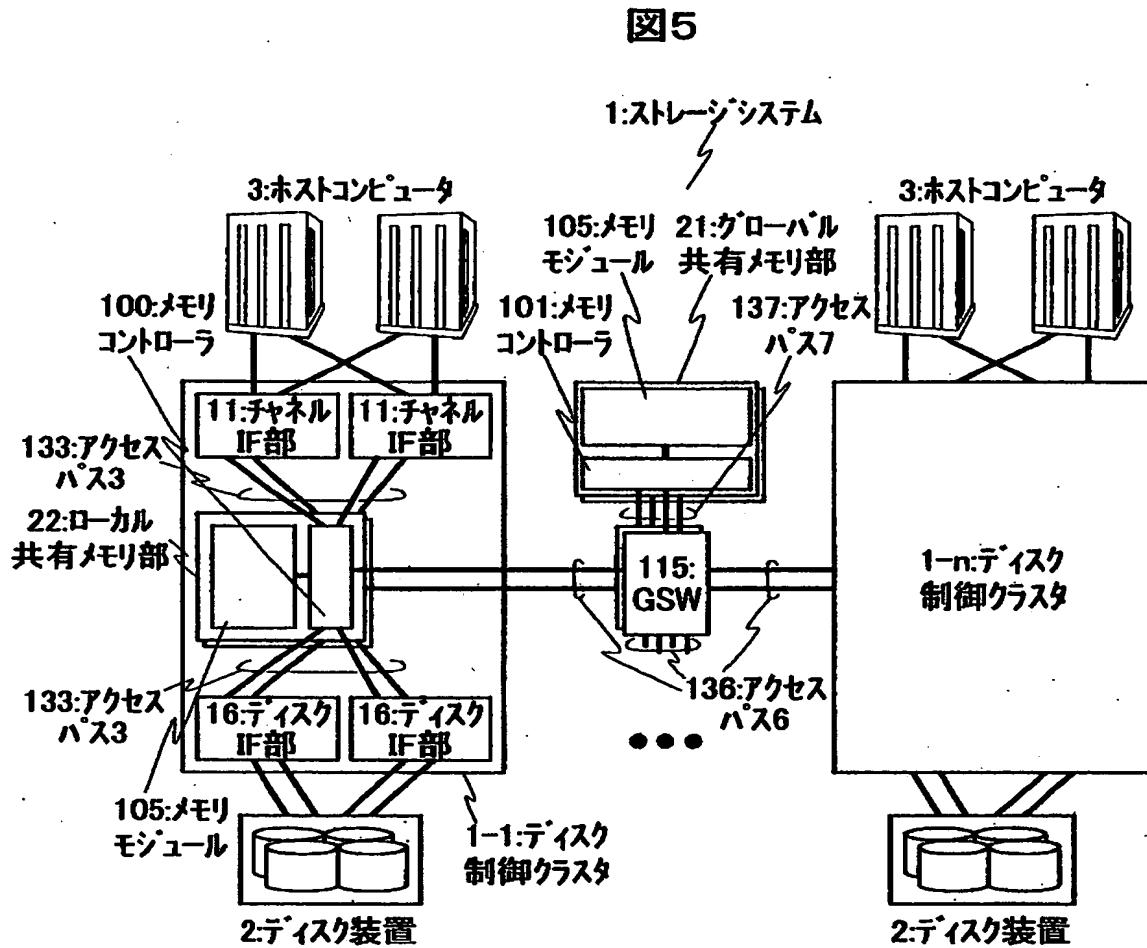


【図4】

図4

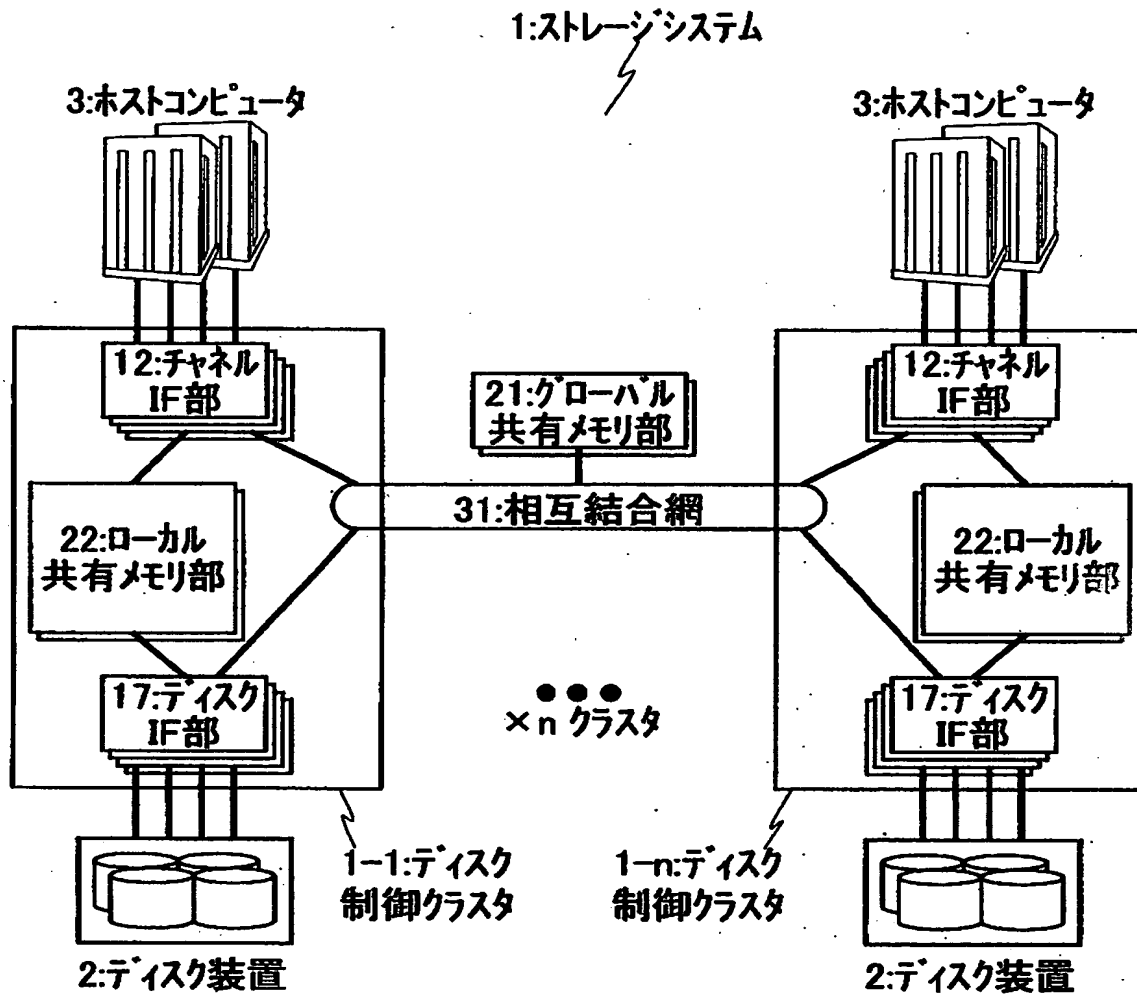


【図 5】



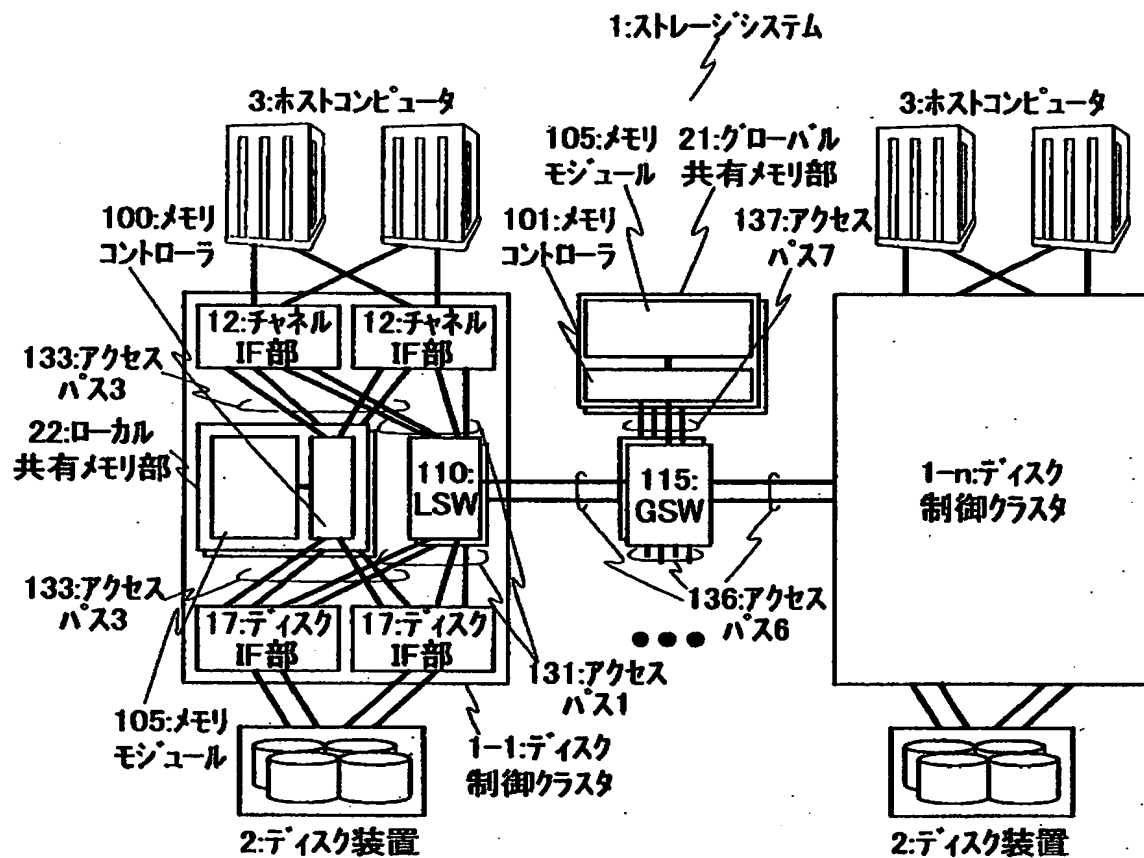
【図6】

図6

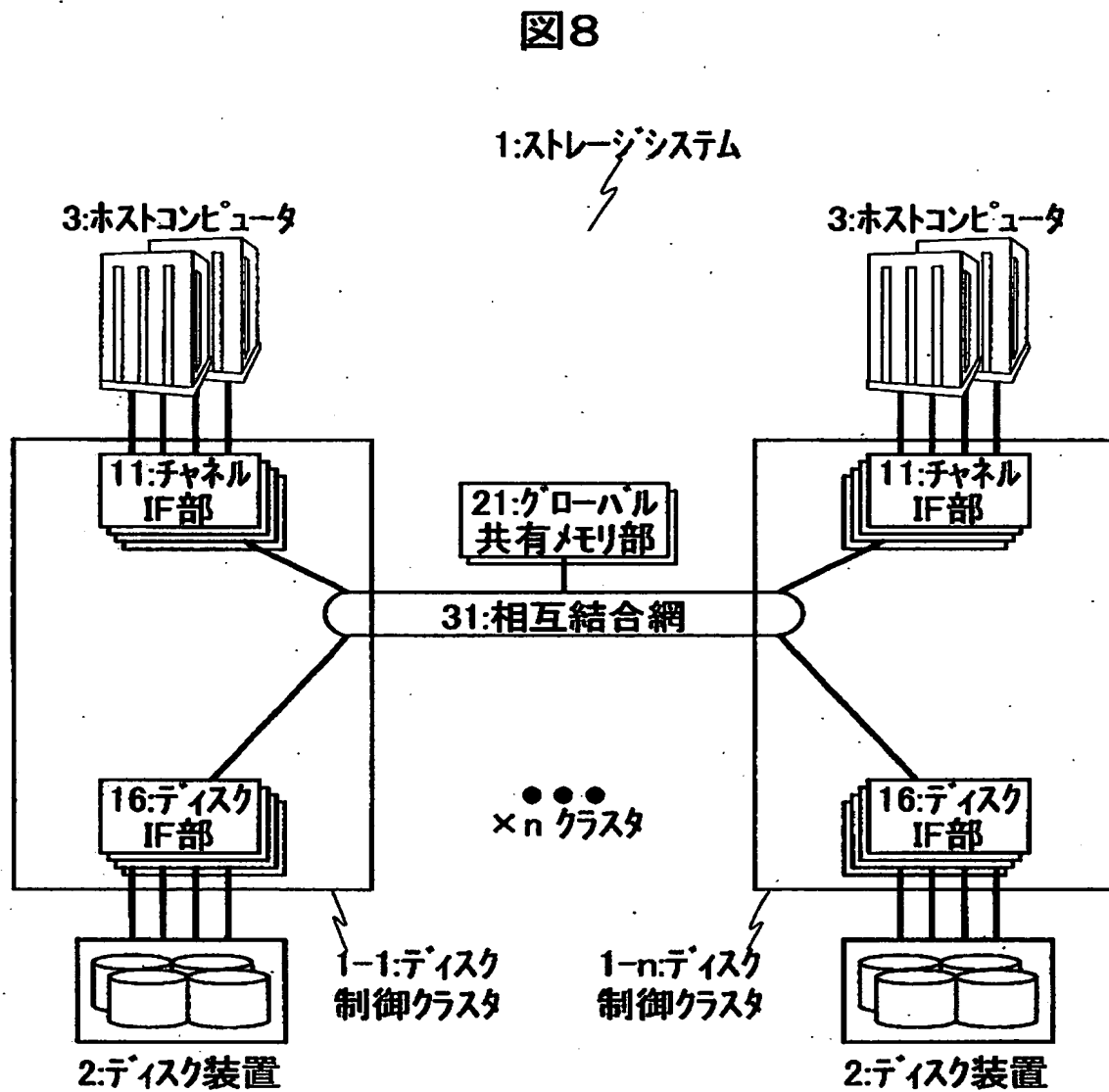


【図 7】

図 7

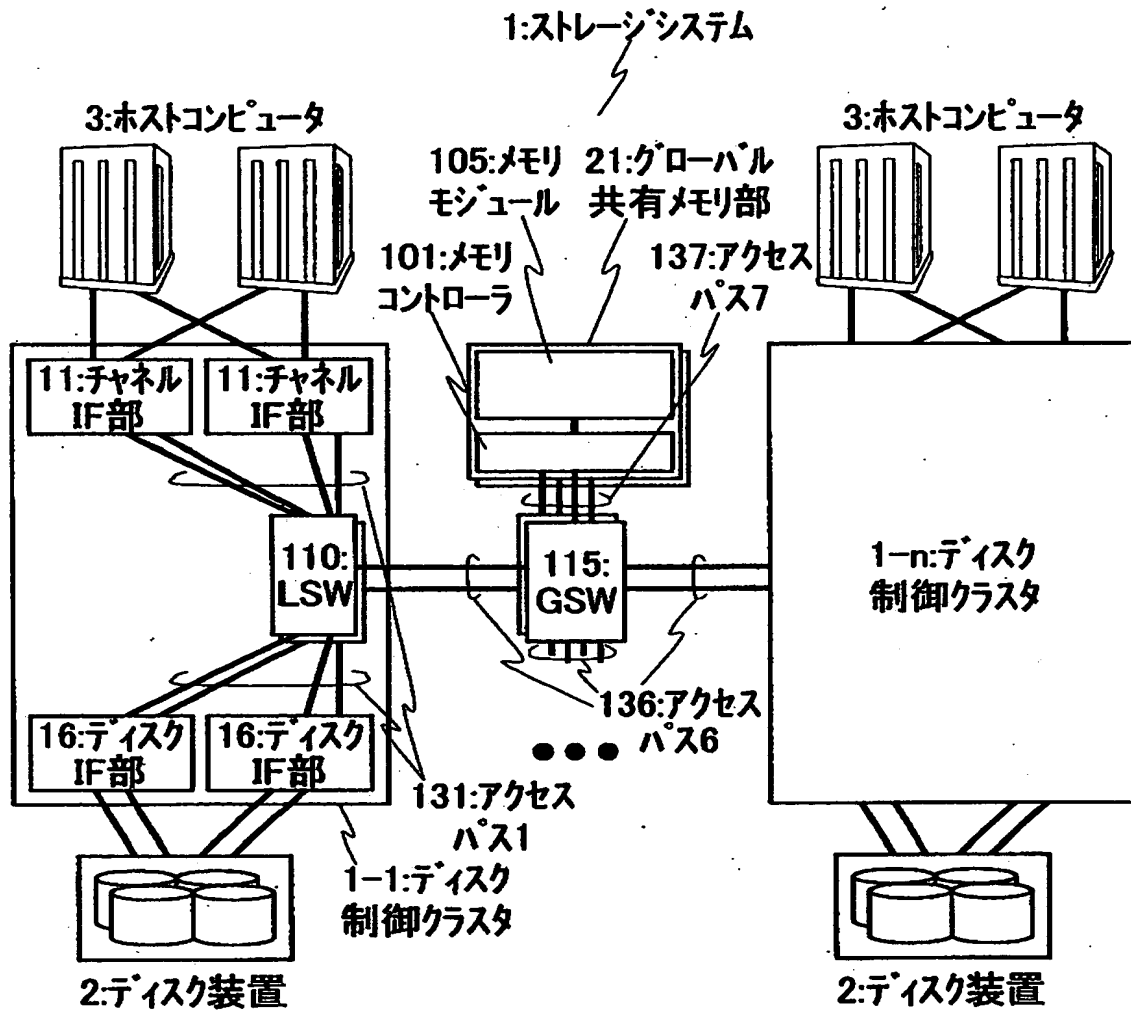


【図 8】



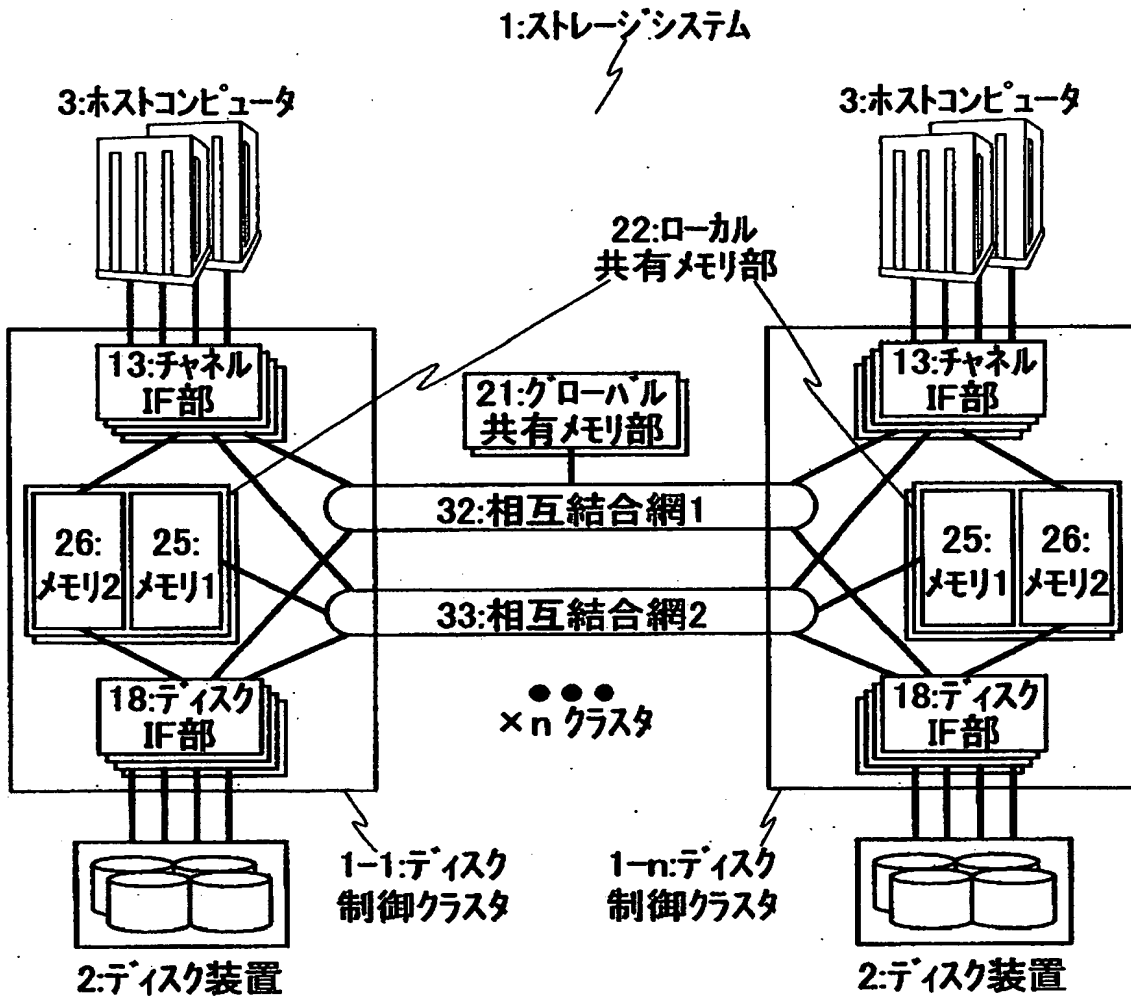
【図9】

図9



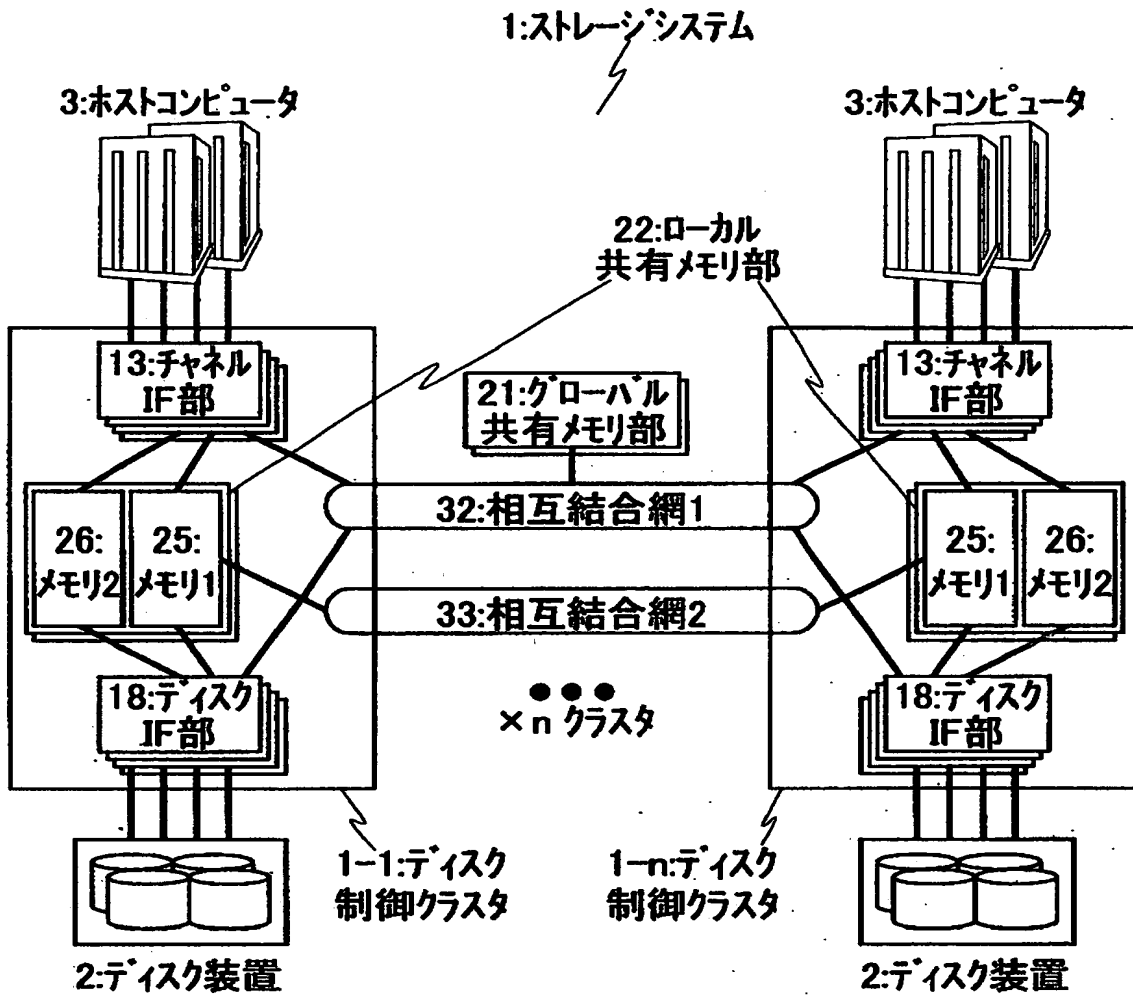
【図10】

図10



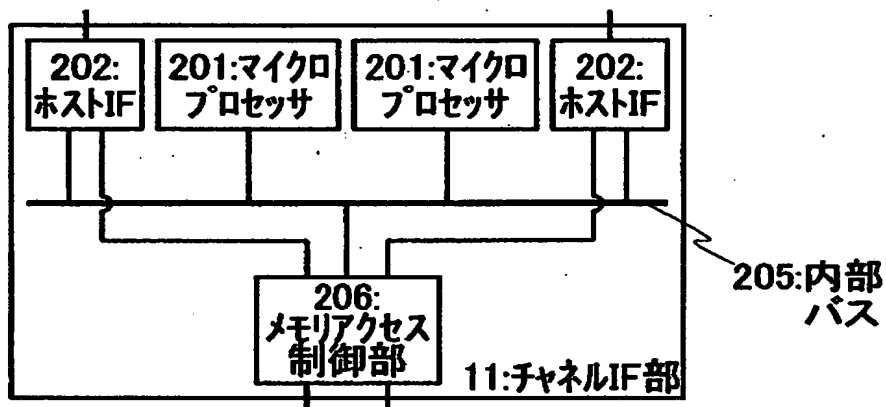
【図11】

図11



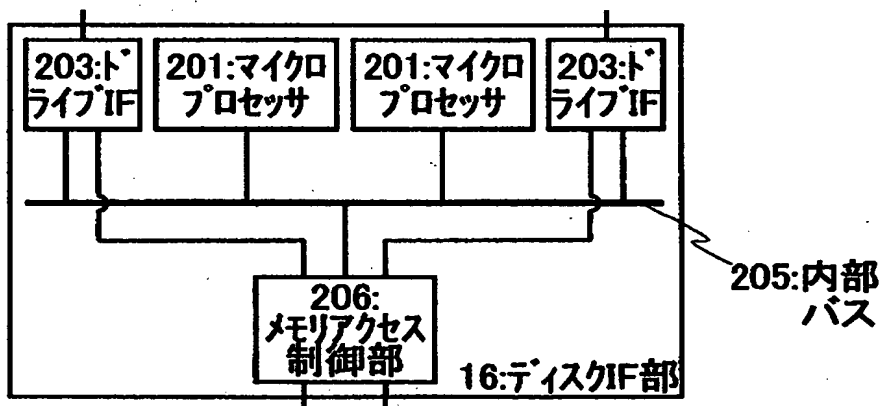
【図 12】

図12



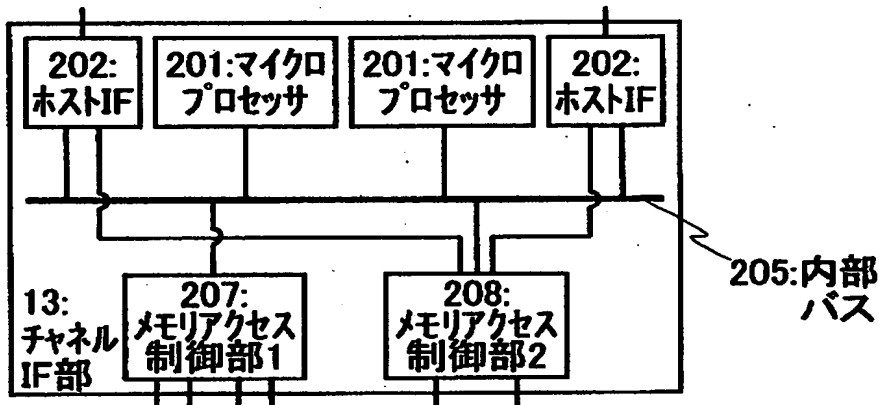
【図 13】

図 13



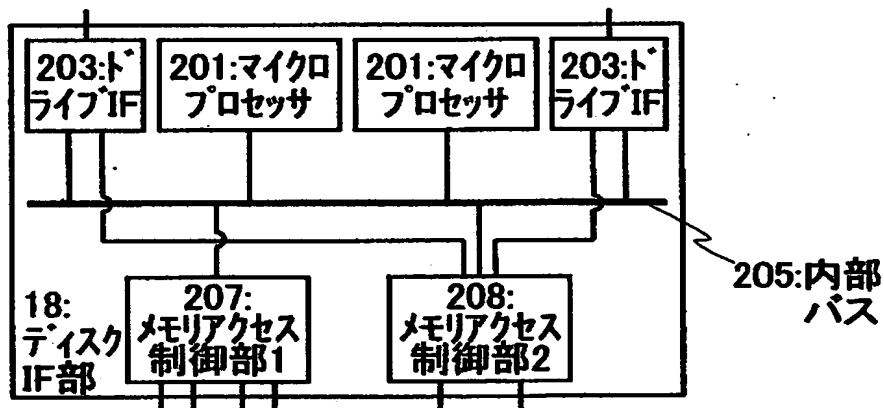
【図14】

図14



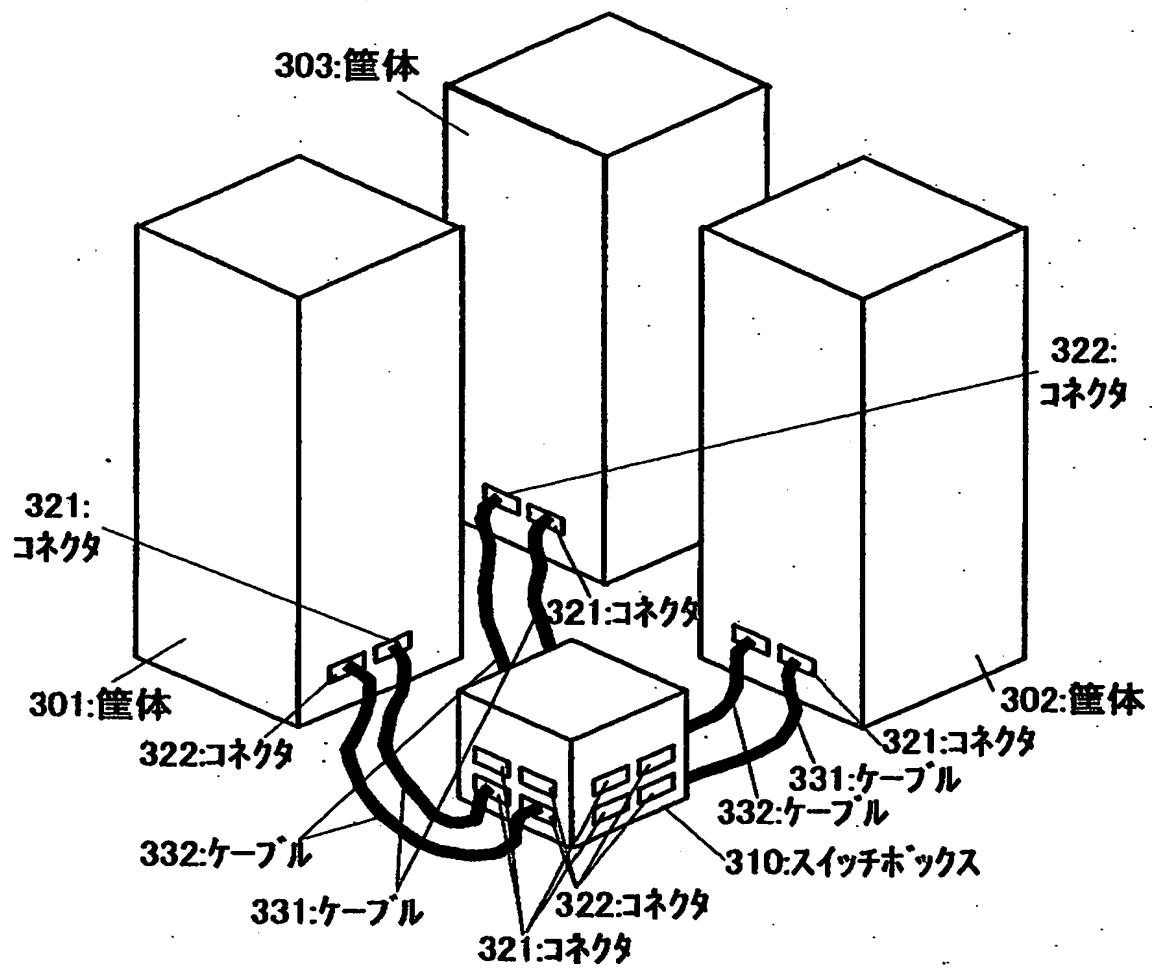
【図15】

図15



【図16】

図16



【図 17】

図 17

(増設前)

GSWポート番号	クラスタ番号
0	0
1	1
2	3
3	2
4	未接続
5	4
6	未接続
7	未接続

400:GSWポート-クラスタ対応テーブル

(増設後)

GSWポート番号	クラスタ番号
0	0
1	1
2	3
3	2
4	5
5	4
6	未接続
7	未接続

400:GSWポート-クラスタ対応テーブル

【図 18】

図 18

(増設前)

クラス番号	論理ボリューム番号
0	0~4095
1	4096~6143
2	12288~16383
3	6144~12287
4	16384~16639
5	未実装
6	未実装
7	未実装

405:クラス-論理ボリューム対応テーブル

(増設後)

クラス番号	論理ボリューム番号
0	0~4095
1	4096~6143
2	12288~16383
3	6144~12287
4	16384~16639
5	16640~20735
6	未実装
7	未実装

405:クラス-論理ボリューム対応テーブル

【書類名】 要約書

【要約】

【課題】 小規模から超大規模な構成まで、同一の高信頼・高性能なアーキテクチャで対応可能な、スケーラビリティのある構成のストレージシステムの提供。

【解決手段】 図はストレージシステム1の構成例を示し、LSW110はローカルスイッチであり、GSW115はグローバルスイッチであり、21はグローバル共有メモリである。例えば、ホスト3からディスク制御クラスタ1-1にデータの読出し要求があった場合、チャンネルIF部11は、LSW110を介してローカル共有メモリ部22にアクセスし、データがディスク制御クラスタ1-1内に有れば、ローカル共有メモリまたはディスク装置2からデータを読出し、ホストに送り、データがディスク制御クラスタ1-1内に無ければ、グローバル共有メモリ部21にアクセスし、要求データが格納されているディスク制御クラスタを調べ、要求データの格納されているディスク制御クラスタから要求データを取得し、ホストに送る。

【選択図】 図3

特2001-294048

出 願 人 履 歴 情 報

識別番号 [000005108]

1. 変更年月日	1990年 8月31日
[変更理由]	新規登録
住 所	東京都千代田区神田駿河台4丁目6番地
氏 名	株式会社日立製作所

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.